2020-08

A reference machine learning model for prediction of cholera epidemics based-on seasonal weather changes linkages in Tanzania

Leo, Judith

NM-AIST

A REFERENCE MACHINE LEARNING MODEL FOR PREDICTION OF CHOLERA EPIDEMICS BASED-ON SEASONAL WEATHER CHANGES LINKAGES IN TANZANIA

Judith Leo

A Thesis Submitted in Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

August, 2020

ABSTRACT

The Cholera epidemic remains a public threat throughout history, affecting vulnerable populations living with unreliable water and sub-standard sanitary conditions. Studies have observed that the occurrence of cholera has also, strong linkage with seasonal weather patterns. Over the past decades, there have been great achievements in developing cholera epidemic models which have focused on using mathematical techniques. However, most existing prediction systems have some challenges such as lack of flexibility, not user friendly, in-effective and also, lack integration of essential weather variables. In addition, the use of advanced technology such as machine learning (ML) have not been explicitly deployed in modeling cholera epidemics in developing countries including Tanzania; due to the challenges that come with its datasets such as missing-information, data-inconsistency, imbalance-class and other uncertainties.

The aim of this work was to overcome and complement the existing challenges of cholera epidemic models by taking the advantages of ML techniques. Hence, by developing an ML model that is capable of predicting cholera epidemic outbreaks based-on seasonal weather changes linkages in Tanzania. Secondary datasets from Tanzania Meteorological Agency (TMA), the Ministry of Health and Social Welfare, and Dar es Salaam Water and Sewerage Authority (DAWASCO) were used. Then, Adaptive Synthetic Sampling Approach (ADASYN) and Principal Component Analysis (PCA) were applied to restore sampling balance and dimensions of the dataset. In order to determine which ML algorithms were best able to predict (yes/no) whether cholera epidemic would occur given the weather variables, ten classification algorithms were evaluated using F1-score, sensitivity and balanced-accuracy metrics. The Friedman-test was then used to determine whether the performance of the models was statistically significant. Results showed that Random Forest, Bagging, and ExtraTree classifiers had the best performance, with 74%, 74.1% and 71.9% accuracy respectively. The ensemble method of model fine-tuning was then applied in order to obtain one model from the three, and an overall accuracy of 78.5% was achieved. Lastly, a model evaluation process was performed on the selected final model. The model validation process involved four processes: The first evaluation process re-ran the final model using the same dataset but without the weather variables; which resulted into confirming that the model with weather variables to have higher performance compared to the model without the weather variable. The second evaluation process re-ran the model-development procedure using

datasets from Tanga and Songwe regions in order to illustrate on how the adaptive reference model can be referenced by other researchers. The third and fourth model evaluation involved mixed-design approach of quantitative and qualitative methods using focus group discussions and interviewer-administered questionnaires with 500 and 20 stakeholders (including; medical officers, epidemiological analysts, nurses, environmental experts, ICT experts and cholera patients) respectively. The results of the third evaluation process proved that 90% of the responses agreed that, the developed model is robust and appropriate to work in least developing countries towards effective prediction of cholera epidemics. Whereas, the results of the fourth evaluation process proved also that cholera ML model is better in terms of their usability, expandability and computational complexity compared to the cholera statistical models.

Overall, the study improved our understanding of the significant roles of ML strategies in health-care data. However, the study could not be treated as a time series problem due to data collection bias such as data-inconsistency in terms of time. The study recommends a review of health-care systems in order to facilitate quality data collection and further deployment of ML techniques in the health sector in Tanzania.

DECLARATION

I, Judith Leo do hereby declare to the Senate of the Nelson Mandela African Institution of Science and Technology (NM-AIST) that this dissertation is my own original work and that it has not been submitted for consideration of a similar degree award in any other University.

Judith Leo

Candidate Name	**Signature**	**Date**

The above declaration is confirmed

Dr. Kisangiri Michael

Supervisor 1	**Signature**	**Date**

Dr. Edith Luhanga

Supervisor 2	**Signature**	**Date**

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read and hereby recommend for submission to the Nelson Mandela African Institution of Science and Technology (NM-AIST) a dissertation titled: "*A Reference Machine Learning Model for Prediction of Cholera Epidemics based on Seasonal Weather Changes Linkages in Tanzania*" in fulfilment of the requirements for the degree of Doctor of Philosophy in Information and Communication Science and Engineering at NM-AIST.

Dr. Kisangiri Michael

Supervisor 1	**Signature**	**Date**

Dr. Edith Luhanga

Supervisor 2	**Signature**	**Date**

ACKNOWLEDGEMENTS

First and foremost, I thank Almighty God for His love and the gift of life. The accomplishment of this study is truly by His Grace.

Secondly, I extend my special thanks to the Nelson Mandela African Institution of Science and Technology (NM-AIST) for granting me the opportunity to pursue my PhD degree and African Development Bank (AfDB) for funding my studies.

Thirdly, I am sincerely gratitude to my supervisors, Dr. Kisangiri Michael (NM-AIST) and Dr. Edith Luhanga (NM-AIST); my mentors Prof. Jorge Gomez, Dr. Patrick Boily and Dr. Stefan Oehmcke, for their tireless support and guidance in the making of this study. Their encouragement, know-how and mentorship have tremendously helped in moulding me into a research scientist.

Fourthly, I would like to appreciate Dr. Dina Machuve, Dr. Anael Sam, Prof. Dmitry Kuznetsov, Prof. Verdiana Massanja and Dr. Musa Ally Dida for their useful comments and constructive criticism during graduate seminars and defense. I also am grateful to hospitals in Arusha, Kilimanjaro and Dar es Salaam, Tanzania Meteorological Agency and the Ministry of Health and Social Welfare for their willingness and support in obtaining data and whichever information I needed to accomplish the study.

Finally, I am forever grateful to my husband, Surgeon Andrea Michael Kileo, my sons Ryan, Rylan and Arvin; my parents, Mr. Leo Herman Lwekamwa and Mrs. Leonida Leo; my brothers, Amon, Ruga, Wilson, Victor and Prince; my three sisters, Edna, Lilian, Gloria and Team MO; as well as my beloved friend Dr. Ruthbetha Kateule for their moral and material support, faith and prayers throughout the entire period of my study.

To all of you, I am truly thankful and may God bless you!

DEDICATION

I dedicate this dissertation to my parents, husband Sur. Andrea Michael Kileo and my sons Ryan and Rylan. You have been very supportive to me in completing my PhD degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS

>	Greater than
$	US dollar
%	Percentage
.csv	Comma Separated Variable
.xls	Microsoft Office Excel of 2013
◦C	Degree Celsius
ADASYN	Adaptive Synthetic Sampling Approach
ANOVA	Analysis of Variances
ARPM	Adaptive Reference Predictive Model
CFR	Case Fatality Rate
cvs	Comma-separated values document
DAWASCO	Dar es Salaam Water and Sewerage Corporation
DL	District Level
DOS	Date on Set
DT	Decision Tree
exl	Excel document
Fig	Figure
FN	False Negative
FP	False Positive
GIS	Geographical Information System
GPS	Global Positioning System
GR	Green Rectangles
GTFCC	Global Task Force on Cholera Control

HEMISs	Health and Environmental Management Information Systems
HIS	Health Information System
HMIS	Health Management Information System
HTML	HyperText Mark-up Language
ICT	Information Communication Technology
IEEE	Institute of Electrical and Electronics Engineers
KCMC	Kilimanjaro Christian Medical Centre
K-NN	K Nearest Neighbours
Lab	Laboratory
LR	Linear Regression
max	Maximum
min	Minimum
ML	Machine Learning
MLP	Multi-Layer Perceptron
mm	Millimetre
NGOs	Non-Government Organizations
NM-AIST	Nelson Mandela African Institution of Science and Technology
No, #, or N	Number
PCA	Principal Component Analysis
RC	Red Circles
RF	Random Forest
SDG	Sustainable Development Goals
SIR	Susceptible Infected Recovered

SMOTE	Synthetic Minority Oversampling Technique
SPSS	Statistical Package for Social Science Software
SSA	Sub-Saharan Africa
SVM	Support Vector Machine
Temp_max	Maximum Temperature
Temp_mean	Mean Temperature
Temp_min	Minimum Temperature
Temp_range	Temperature Range
TMA	Tanzania Meteorological Agency
TP	True Positive
V. cholerae	*Vibrio cholerae*
WASH	Water, Sanitation and Hygiene
WHO	World Health Organization
Wind_Dir	Wind Direction
Wind_Spd	Wind Speed
WL	Ward Level

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Cholera is an acute epidemic infectious disease caused by *Vibrio cholerae (V. cholerae*) bacteria (Li, Wang & Di, 2017). The bacteria typically live in salty-warm waters along the coast. Human beings contract *V. cholerae* through ingesting liquids or foods contaminated with the bacteria (Xu *et al*., 1982). The disease remains notorious and threatening to human societies throughout history, due to the extraordinary scale of deaths and damage it has brought over the years (Mandal, Mandal & Pal, 2011).

Cholera is transmitted from one person or place to another through the faecal-oral route of contaminated food or water caused by poor sanitation (Symington, 2011). Food transmission can occur when people harvest seafood such as oysters and shellfish in the waters infected with *V. cholerae.* People infected with cholera often have diarrhea and hence, disease transmission may occur if this diarrhea contaminates water used by other people (Miller, Feachem & Drasar, 1985). A single diarrheal incident can cause a one million increase in numbers of *V. cholerae* in the environment through waterways, groundwater and drinking water supplies. Normally, the transmission of cholera directly from person to person is very rare (Fung, 2014).

Vibrio cholerae can also exist outside the human body in natural water sources, either by itself or in association with phytoplankton, zooplankton and biotic and abiotic detritus. Hence, drinking such water can also result in cholera disease, even without prior contamination through faecal matter (Joachim & Karl, 2002). In addition, there are several virulence factors which can easily contribute to the pathogenicity of the *V. cholerae* to easily infect and cause symptoms to the hosts (Novais *et al*., 1999). These virulence factors include toxin co-regulated pilus, cholera toxin and motility (Siettos & Russo, 2013). Furthermore, in our rapidly changing environment, it has been reported by several researchers that the transmission and infection of cholera epidemics are greatly influenced by seasonal weather variation (Lugomela *et al.,* 2015). This is because the dynamics of weather patterns dictate the infection and transmission rate of cholera disease. As they affect natural demographic behaviour of population involved and also, influences almost all variables involved in the growth of *V. cholerae.* Moreover, the fluctuation of weather variables such as temperature,

rainfall, humidity and wind, is also regarded as the core factor that causes re-emergence of cholera outbreak cycles and its variability from small to large scales (Constantin *et al*., 2008).

1.1.1 Scale and Effects of Cholera Epidemics

In early days, the cause of cholera was not known. This caused massive panic and death of millions of people across the world wherever an outbreak occurred (Sinha *et al*., 2002). Cholera epidemic/outbreak refers to a widespread occurrence of the disease in a community at a particular time; whereas a cholera epidemic that lasts for many years or even a few decades at a time and that spreads to many countries and across continents and oceans is known as a cholera pandemic (Karaolis *et al*., 1995). Literature shows that, there have been a total of seven cholera pandemics (Lessler *et al*., 2018). The first cholera pandemic occurred from 1817 to 1824 in India and spread to Southeast and Central Asia, the Middle East, China and Russia, leaving hundreds and thousands of people dead (Lan & Reeves, 2006). The second pandemic occurred in 1826 to 1837 in India and spread to western Asia, Europe, Great Britain and the Americas, as well as East of China and Japan. It caused more deaths, more quickly than any other epidemic disease in the 19th century (Chan, Tuite & Fisman, 2013). The third pandemic also, caused the highest fatalities in the 19th century (Azizi & Azizi, 2010). It originated in India and spread far beyond its borders to Russia and Great Britain. Researchers at the University of California in Los Angeles believe that the third cholera pandemic started as early as 1837 and lasted until 1863. In 1853 to 1854, the pandemic caused 23 000 deaths in Great Britain and over 10 000 deaths in London. As a result of the August 1854 cholera outbreak in London, the physician John Snow identified contaminated water as the means of transmission of the disease. He mapped a cluster of cholera cases near a water pump in one neighbourhood. His breakthrough led to the control of cholera epidemics around the world in the 19th century (Azizi & Azizi, 2010).

However, there were other cholera pandemics after John Snow's breakthrough, such as the fourth cholera pandemic which began in 1863 and ended in 1875, the fifth pandemic (1881 to 1896), the sixth (from 1899 to 1923) and the seventh (from 1961 to the 1970s) (Glass & Black, 2003) (Karaolis *et al*., 1995). During the fourth pandemic, cholera spread throughout the Middle East and was carried to Russia, Europe, North America and reached North Africa where it spread to Sub-Saharan Africa (SSA), killing 70 000 in Zanzibar - Tanzania in 1869 (Towner *et al*., 1980). Generally, cholera is still prevalent in SSA and Asia areas with

inadequate sanitation, poor food, water hygiene, in-effective early warning systems and also, remains a major global public health problem (Weill *et al*., 2017), as indicated in Fig. 1.

It is still estimated globally that every year, there are roughly 1.3 to 4.0 million cases and 21 000 to 143 000 deaths worldwide due to cholera. In Tanzania, recent outbreaks include the August 2015 to April 2016 outbreak in Dar es Salaam region in Tanzania which led to a total number of 14 608 cases and 228 death incidences (Narra *et al*., 2017); this case can be statistically shown in Fig. 2.

Researchers believe that timely detection of cases of cholera is essential in order to prevent and limit the severity and duration of the outbreak (Kuhn *et al.*, 2004; World Health Organization, 2016). Therefore, there is a dire need to implement measures for early prediction mechanisms which could help public health officials to prepare for containment before cholera epidemic and pandemic occur (Fung, 2014; Burnett *et al*., 2016).

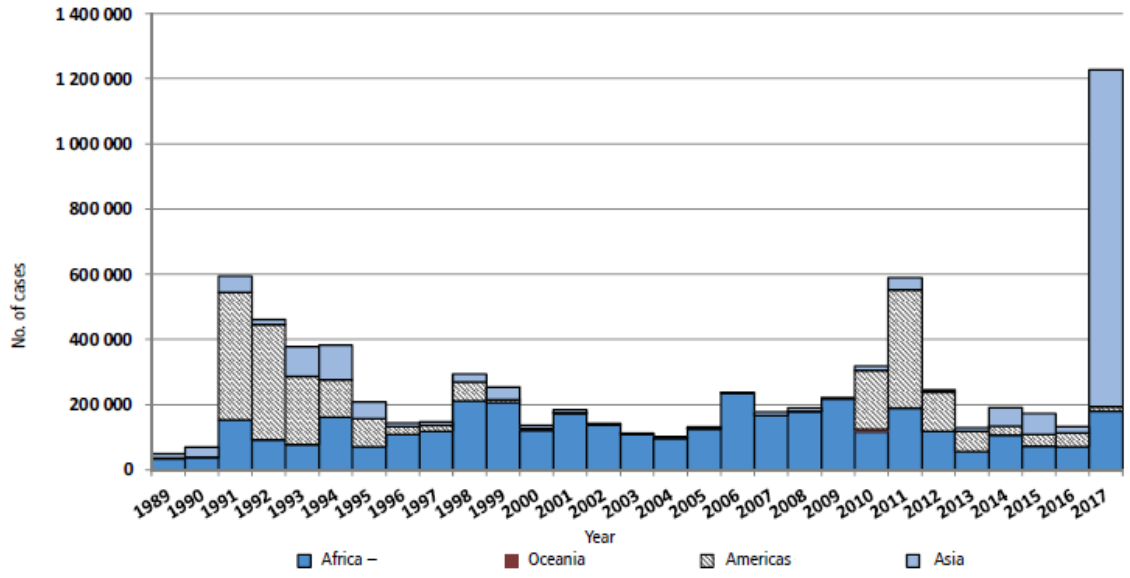


Figure 1: Cholera Cases Reported by WHO by Year and by Continent from 1989-2017 (WHO, 2018)

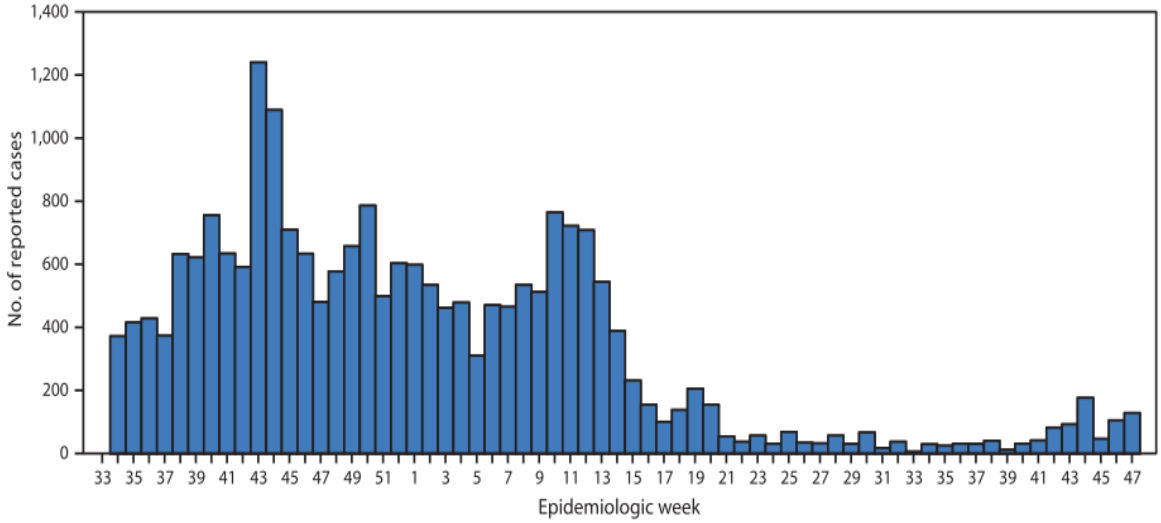


Figure 2: Number of Reported Cholera Cases in Tanzania- August 15, 2015, to November 26, 2016 (Narra *et al*., 2017)

1.1.2 Factors Affecting Transmission and Infection of Cholera in Developed and Least Developing Countries

Cholera-endemic countries are countries with a cholera outbreak within the last three (3) years. These countries include: India, Ethiopia, Nigeria, Haiti, the Democratic Republic of Congo, Tanzania, Kenya and Bangladesh (Mukhopadhyay *et al.*, 2015). Cholera is endemic in these countries because of different modes of transmission and infection which include: social issues such as poor hygiene and sanitary conditions as well as technical issues such as ineffective cholera management systems for analysis and prediction and environmental issues such as; climate change (Legros, 2018).

It has been noted by researchers that developed and least developing nations have somehow different modes of cholera transmission. Most of the cholera cases in developed countries are transmitted through contaminated food, whereas in least developing countries, it is more often through contaminated water (Mandal, Mandal & Pal, 2011; Huq & Colwell, 1996). This is due to the fact that developed countries have good sewage systems in place (Descamps *et al.*, 2012) and thus water is rarely contaminated. In addition, most least developing countries fall in regions with Tropical and Mediterranean climates whereas developed nations such as North America and German are in the colder temperate regions, where environmental conditions are less favourable for *V. cholerae* growth (Leotta *et al*., 2006) and (Jutla *et al*., 2013). Developed countries are also affected through various human migration. This fact is

shown clearly on the world map in Fig. 3 where countries with reported cholera cases in the red pattern import cholera to countries with the yellow pattern.

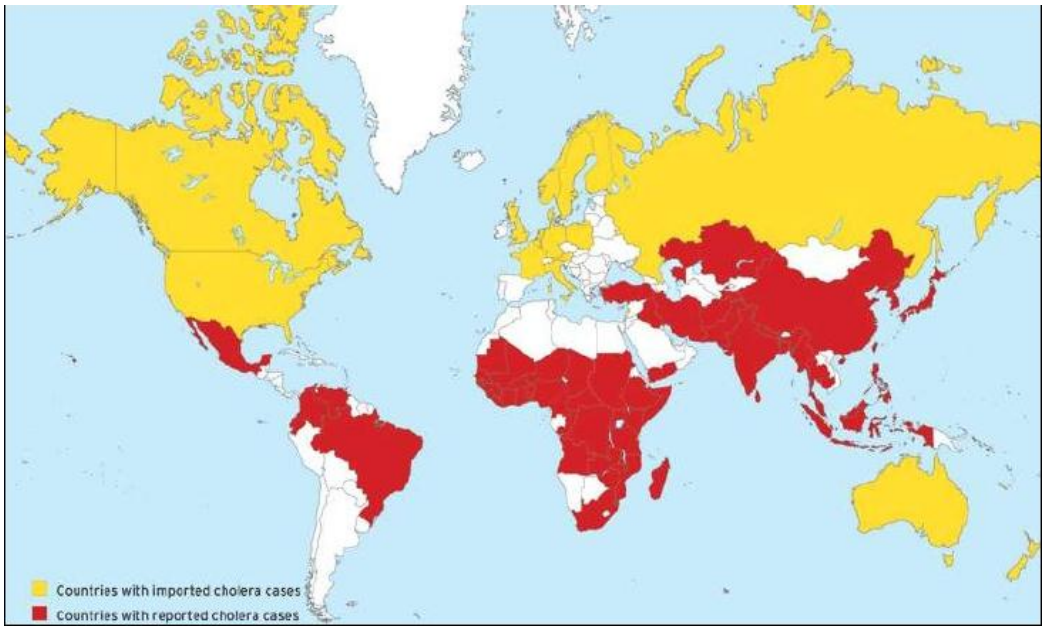


Figure 3: Cholera Epidemics Imported from One Continent to Another (WHO, 2011)

1.1.3 The Role of ICTs in Controlling Cholera Epidemics

The Global Task Force for Cholera Control (GTFCC) proposed a global strategy for cholera control at the country level, to reduce the number of cases by 90% by 2030 (Legros, 2018). Two of the proposed strategies is the development of early detection and rapid response mechanisms and implementation of Multi-Sectoral Approach to contain cholera outbreaks and the report also highlighted the positive impact that information and communication technologies (ICTs) can have in early detection and multi-sectoral approach.

In many countries, disease surveillance systems such as Integrated Disease Surveillance and Response (IDSR), District Health Information System 2 (DHIS2) and traditional statistical techniques such as mathematical models have been the main tools used for tracking outbreaks of water-borne, vector-borne and vaccine-preventable diseases (Das *et al*., 2019). The systems have enabled the management of epidemics such as cholera to some extent; however, several challenges remain. First, most of the existing systems and models are not capable of handling large amounts of data with complex nature such as missing information, imbalanced data, and other uncertainties; Examples of these challenges include essential dataset for disease prediction in different format and platform; due to the rapid global climatic change, epidemiologists are not able to predict the disease outcome through their experiences and

also, Microsoft Excel document is not able to easily predict the relationship of two variables for a dataset with more than 5000 values which has data inconsistency and imbalance limitations (Karimuribo *et al*., 2011; Yesserie, 2015). Challenges such as missing information may require systemic changes such as organisational workflow changes and personnel skills training (Pascoe *et al*., 2012). Second, the data collected and stored in most of the existing systems do not include weather variables and other environmental variables necessary for the analysis and prediction of cholera epidemics (Dent, 2012) and (Leo, Michael & Kalegele, 2015). As the global burden of cholera epidemics from the seasonal weather variables and other environmental factors such as poor sanitary condition is expected to increase over time with a rapid increase of epidemic size (Trærup, Ortiz & Markandya, 2010), it is important that these factors be included in any surveillance and response systems. Finally, existing systems still require a number of tasks to be performed manually as reported by Pascoe *et al*. (2012) and Renggli *et al*. (2018). With the increase of number of epidemic predictors from the effect of climate change and the existence of limited number of the workforce in the least developing countries' health-sectors, for instance in Tanzania's hospitals, the use of existing manual mechanisms and ineffective tools for managing cholera epidemics is a great challenge (Kwesigabo *et al*., 2012). Manual analysis is also time-consuming and can prevent timely prediction and response to disease outbreaks (Kuhn *et al*., 2005). Overall, the lack of flexibility and user friendliness of existing systems makes them challenging for use in early prediction of disease outbreaks (Renggli *et al*., 2018). Hence, there is a dire need to make use and take advantages of other advanced and emerging technologies such as Machine Learning (ML) in order to assist existing systems in cholera management, analysis, prediction and control.

Machine learning (ML) is an application of artificial intelligence that provides computer-based systems the ability to automatically learn and improve from experience from datasets without being explicitly programmed (Demšar *et al*., 2013). It has proven to be an effective prediction tool in many domains including the health domain. For example, it has successfully been applied in identifying disease and diagnosis, drug discovery and in predicting various disease outbreaks especially in developed countries settings, due to the availability of quality data and long-time funded programs for such activities (Peng *et al*., 2015). Despite its potential to resolve many of the challenges faced with IDSR, DHIS2 and traditional statistical epidemic models in predicting cholera outbreaks, it has been underutilised in least developing countries including Tanzania. This situation has been

influenced by the risk to patient privacy of data acquisition, risk of injuries to patients from ML system errors, lack of quality datasets (Pascoe *et al*., 2012; Kwesigabo *et al*., 2012), as well as a lack of good knowledge on emerging and advanced ML strategies which can work innovatively with any dataset toward fulfilling the intended goal (Raschka, 2018; Nejatimoghadam, 2018; Domingos, 2012; Chao, 2011).

This research study therefore, researches into how well ML models can analyse and predict cholera epidemics given data available in IDSR and DHIS2, as well as weather variables available in environmental management information systems (EMIS) of least developing countries.

1.2 Statement of the Problem

It is estimated that every year, around 21 000 to 143 000 people die due to cholera. Timely detection of cases of cholera can limit the spread, severity and duration of cholera outbreak (Kuhn *et al*., 2004) and (World Health Organization, 2016). Providing early detection tools to public health officials and the community is therefore necessary to prevent cholera epidemics and pandemics (Fung, 2014; Burnett *et al*., 2016).

Many least developing countries including Tanzania now use IDSR and HMIS to collect and store data of infectious diseases such as cholera and they rely on the use of mathematical models and other traditional statistical techniques for prediction of disease outbreaks. However, these techniques are in infective, complex to understand and extend for non-mathematicians and also, do not include environmental variables such seasonal weather variables which researchers posit are necessary for accurate cholera epidemic prediction. They are also ill-equipped at dealing with large amounts of dataset with complexities such as missing values and data-imbalance conditions.

Machine learning techniques are adept at providing accurate predictions in various applications such as disease diagnosis, prediction and drug discovery especially with quality data. Despite this, they have been underutilised in predicting cholera epidemics in least developing countries and it is thus unknown how effective they are in supporting early predictions. This work therefore aimed at taking advantage of the prediction strength, adaptability nature, and innovative features of ML techniques to develop an ML model for prediction of cholera epidemics based on seasonal weather changes linkages in Tanzania, and comparing its performance (accuracy, computational resources needed, user-friendliness) to

that of existing cholera mathematical models, in order to determine its efficacy for effective cholera epidemic prediction.

1.3 Rationale of the Study

In 2017, WHO's study reported that there is a lack of effective and reliable tracking system for early warning prediction, control, and prevention of epidemics (Darcy *et al*., 2015; WHO, 2018). Moreover, statistics show that health and environmental workers, especially the epidemiologists, statistician and ecologists fill only 35% of positions available in most least developing countries, leaving most countries with a severe human resource crisis in the health and environmental sectors (Kwesigabo *et al*., 2012; Manzi *et al*., 2012; WHO, 2015). With the growing number of healthcare data and increase of epidemics size, this figure (35%) of specialists is too small to support the use of ineffective, traditional prediction, manual-based systems and different sources of information to track environmental determinants for epidemics control and management. No wonder the occurrences and recurrences of epidemic outbreaks, such as cholera are affecting most of the least developing countries and the world at large (Castellazzo *et al*., 2012; Impouma *et al*., 2007). In this case, there is a need of facilitating easy proper availability and development of effective mechanism to enable the integration of all necessary datasets from environmental determinants and cholera variables in order to enhance effective management of cholera epidemics.

1.4 Objectives

1.4.1 Main Objective

The main objective of the study was to develop an effective adaptive reference ML model for prediction of cholera epidemics based on variables observed in hospital patients as well as the seasonal weather changes linkages in Tanzania.

1.4.2 Specific Objectives

The following specific objectives were used as a guide towards achieving the main objective:

(i) Review climatic conditions and other requirements that are important predictors for outbreak of cholera epidemics.

(ii) Develop and validate a system for linking health-care and seasonal weather data in order to automatically build cholera datasets integrated with important climatic variables.

(iii) Develop a Reference ML (RML) model for prediction of cholera epidemic based on seasonal weather changes linkages in Tanzania.

(iv) Evaluate the performance of the developed ML model.

1.5 Research Questions

Respectively mapping the research specific objectives, the following research questions were addressed in order to successfully achieve the research main objective:

(i) What are the appropriate model requirements for designing an effective RML model which will be used to predict the occurrence of cholera epidemics using seasonal weather changes in Tanzania's settings?

(ii) What feature-designs are most important for developing a climatology-aware HIS? Does the developed system work as expected?

(iii) What feature-designs are most important for developing an effective RML model for prediction of cholera epidemics and weather changes linkages in Tanzania's settings?

(iv) Is the developed model robust for prediction?

1.6 Significance of the Study

The study is significant in facilitating the development of cholera predictive models through the use of the proposed adaptive reference ML model. Further, the study offers the following advantages:

(i) The model can effectively predict of cholera epidemics, thereby enabling the public to well prepare against the outcomes of seasonal weather changes in relation to cholera epidemics.

(ii) The study facilitates research development in integrated and interoperable healthcare systems such as Health and Environmental Management Information Systems (HEMISs).

(iii) The study offers a review of data collection systems in the health sectors.

(iv) The study facilitates research development in the health sectors using ML in least developing countries' settings.

(v) Lastly, the model will enable the policymakers and other related stakeholders in making effective decisions based on available data in the health sector.

1.7 Delineation of the Study

1.7.1 Proposed General Features of the Proposed Model

The conceptual design of the developed model is shown in Fig. 4. The model developed was designed to work on HMIS patient data to predict whether a patient potentially had cholera (yes/no) based on the symptoms data stored in the HMIS as well as weather variables obtained from EMIS data. If multiple patients were found to potentially have cholera (yes prediction), the HMIS would issue a warning about a possible outbreak and provide details of where the patients came from so follow-ups could be made by health officials.

The model was thus trained using the following features: datasets of cholera epidemic cases such as patient's age, sex, home location, date on set, lab result and his social-economic factors, seasonal weather changes (such as temperature, rainfall, wind and humidity) and clean and waste water distribution in region and district formats.

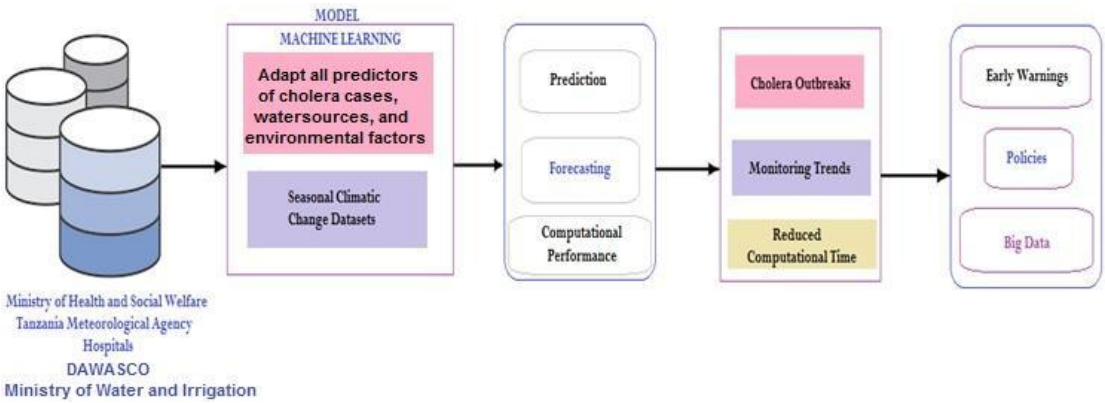


Figure 4: Conceptual Framework of ARML Model

Evaluation results showed that the developed ML model works more effective when environmental variables are used as predictors, can work as a reference model with new datasets from other regions (Tanga and Songwe), provide effective prediction and cholera management and also, works better in terms of in terms of its usability, extensibility (incorporating environmental variables) and computational performance compared to the existing cholera mathematical models.

1.7.2 Scope of the Research Study

The scope of this research study was to develop an adaptive reference ML model for prediction of cholera epidemic outbreaks based on seasonal weather patterns linkages in Tanzania. Therefore, the study was limited to reviewing what other researchers have done on cholera epidemics models, early warning systems for cholera epidemics, linkage of cholera epidemics with environmental factors, cholera management, ML predictive models and techniques of handling complex (such as imbalanced class) and large amount of data. Overall, the scope aimed to get the insight and a deeper understanding of different used techniques and strategies so as to better improve and innovatively combine these strategies towards the development of the proposed ARML model in Tanzania's settings. The study, therefore, started with the analysis and identification of model's requirements. Then, continued with the design, development and evaluation of the proposed model. Finally, it ended with the validation of the developed model.

1.7.3 Limitations of the Research Study

Given the financial demands and time-constraints associated with the development of RML model for prediction of cholera epidemics and seasonal weather changes linkages in such settings as Tanzania's the current study had the following limitations:

(i) The study was done in Tanzania but has only used datasets from Dar es Salaam region to develop the model and only used validating datasets from Tanga and Songwe regions. Therefore, the word Tanzania in the title is represented by the three regions which are; Dar es Salaam, Tanga and Songwe.

(ii) Given that the term 'environmental determinant' is very broad as "it encompass all external factors and conditions that affect people's lives and that from a health perspective, the definition may widely include not only social factors, but also often

times more restricted meaning of chemical, biological and physical agents that impinge on health", the study, thus only focused on explaining the linkages between cholera dynamics and the physical environmental determinants which are weather changes such as rainfall, humidity wind and temperature.

(iii) It adopted only features from cholera epidemics statistical models on how they handle datasets or variables such as population and environmental determinants.

(iv) The hygiene and sanitation level of an area will be represented by the source of drinking water (WasteWater variable) as shown in Table 5.

(v) The results of the research study are subject to or based on the nature of the data collected in Dar es Salaam, Tanga and Songwe regions.

(vi) In this study, Complexity of the data means the challenges of the data such as missing information, data duplicates, imbalance class, data inconsistencies and other errors that will exist in the collected data.

(vii) The Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDEC) will be referred to as Ministry of Health and Social Welfare.

(viii) Data with missing information more than 75% of the total dataset will be reduced / removed.

(ix) The developed climatology-aware has only integrated the weather variables from sensors in customized existing GIS-based HIS.

(x) It should be noted that the occurrence of cholera in an area is represented by the number of cholera patients in that area at the threshold of 1 patient per day as defined by WHO and UNHCR emergence and epidemic standard threshold for cholera epidemics (Strategy, 2018). This is so, because it is believed that a single diarrheal incident can cause a one million increase in numbers of *V. cholerae* in the environment (Fung, 2014). However, this study will take this as guidance and predict according to the nature of the dataset.

(xi) If the model will predict the occurrence of cholera (YES Cholera), while there will be NO cholera, the study will assume that the early awareness and preparedness has assisted in the occurrence of NO cholera.

(xii) It should be noted that the LabResult which is presented by 0 as NO Cholera and 1 as YES Cholera; Meaning that if YES Cholera, the patient has cholera and it also represent the area/ district/ region has cholera disease and vice versa is true.

(xiii) The word weather variables mean weather changes and weather data in the document.

CHAPTER TWO

LITERATURE REVIEW

2.1 Review of Cholera Initiatives and Disease Prediction Models

A review on what other researchers have done in terms of cholera epidemics models, early warning systems for cholera epidemics, linkage of cholera epidemics with other factors such as environmental factors, cholera management, ML predictive models and techniques of handling complex and large amount of data. The chapter also, entails not only the practical challenges in the adoption of ML in healthcare sectors, but also recommendations on proper use of ML functionalities towards developing an effective ML predictive model.

The search of articles involved three web scientific indexing service databases namely Science Direct, IEEE Xplore and Google Scholar. The first step involved searching for the articles by using the created searching query of keywords. Such query keywords include "ML" AND "Disease" OR "Epidemics" OR "Cholera", "Prediction Models" AND "Disease" OR "Epidemics" OR "Cholera" and "Health-sector" AND "Challenges" OR "Least Developing Countries" OR "Tanzania". The second step involved selecting all the papers that were relevant to all targeted areas of the research study. Finally, a total of 50 selected articles were reviewed by analysing addressed problems, their solutions and the study areas.

2.2 Overview of Initiatives to Combat Cholera Epidemics

A number of studies have highlighted the essential factors required for cholera analysis and prediction. Such studies include the study by Codeco (2001), who modelled transmission of cholera using basic Susceptible, Infected and Recovered (SIR) approach, coupled with an aquatic reservoir of *V. cholerae*. Other studies are by Guillaume *et al*. (2008), Escobar *et al*. (2015), Pascual, Bouma and Dobson (2002), Mukhopadhyay (2015), Guillaume, Constantin and Colwell (2009), Lugomela *et al*. (2015), Trærup, Ortiz and Markandya (2011b), Narra *et al*. (2017), Trærup, Ortiz and Markandya (2010b) and Kuiwite *et al*. (2017) on environmental signatures associated with cholera epidemics. Overall, these studies found that ocean waters along with the zooplankton community, weather changes, geo-location, season (date and time), human behaviour and salinity are important factors in the ecology of *V. cholerae* and useful predictors of cholera epidemics. Furthermore, these studies bring some new insights

into cholera epidemiology essentially in developing one platform for collecting all necessary variables necessary for effective cholera epidemic analysis and prediction.

2.2.1 Non-ICT-based Initiatives to End Cholera in Tanzania

There are several non-ICT-based initiatives to counteract the impact of environmental factors in the transmission of diseases including cholera epidemics in Tanzania. Such initiatives include the one health initiative which aims at expanding interdisciplinary collaborations and communications in all aspects of health care for humans, animals, and the environment (Conrad *et al*., 2009). Another initiative is Tanzania's One Health Strategic Plan of 2015-2020 which focuses on five thematic areas namely (Darcy *et al*., 2015): (a) Training, advocacy and communication (b) Preparedness and response (c) Research (d) Disease surveillance, prevention and control and (e) Coordination. Other initiatives include improved water, sanitation and hygiene (WASH) (Boisson *et al*., 2015; Seleman & Bhat, 2016), WHO road map to reach 2030 targets (Molyneux, Savioli & Engels, 2017), Sustainable Development Goals (SDGs) (Fitzpatrick & Engels, 2015) and WHO country cooperation strategy (Yesserie, 2015). In brief, the focus or goals of these initiatives can all be categorized into:

(i) Early Detection and Quick Response to the Control of Cholera Outbreaks

This intervention aims at controlling the outbreaks wherever they may occur through early detection and rapid response mechanisms such as robust community engagement, strengthening early warning surveillance and laboratory capacities and supply readiness (Schroeder, 2013).

(ii) A Targeted Multi-Sectoral Approach to Prevent Cholera Recurrence

The multi-sectoral approach uses a combination of health, finance, and environment sectors to prevent recurrence of cholera on relatively small areas most heavily affected by cholera (Security & Cluster, 2016).

(iii) An Effective Mechanism of Coordination for Technical Support, Advocacy, Resource Mobilization and Partnership at Local and Global Levels

This is done by organizations including Global Task Force on Cholera Control (GTFCC) which are positioned to bring improved WASH through the use of oral cholera vaccines and

offer an effective country-driven platform to support advocacy and communications, fundraising, inter-sectoral coordination and technical assistance (Walldorf *et al*., 2017).

2.2.2 ICT-Based Initiatives to End Cholera Epidemics

In Tanzania, as in least developing countries, there has been a growing effort to strengthen the existing health systems and models (Mghamba *et al*., 2011). These efforts have been mostly concentrated on specific initiatives and programs such as revision of data collection tools for expanded programs of immunization, reproductive and child health services, disease monitoring, analysis, prediction, and evaluation for sexually transmitted and neglected tropical diseases such as HIV-AIDS, malaria and cholera (Kiwanuka, Kimaro & Senyoni, 2015). Study reviews on existing systems such as; OpenMRS and Care2X by (Morrison, 2008; Grad, Miller & Lipsitch, 2012; Pasetto *et al.*, 2017; Koepke *et al.*, 2016; Pascual *et al.*, 2008; Mubangizi, Mwebaze & Quinn, 2009). The results of their reviews provided an insight on the need of design and development of effective health information system which will assist in effective analysis and prediction of different diseases including cholera.

Other systems which were reviewed include an integrated global positioning system (GPS) and geographical information system (GIS) which maps cholera cases in order to investigate the outbreak by using satellite-based recording systems (Mukhopadhyay, 2015). The system potentially acts as an early warning system for cholera in many least developing countries, especially during the start of an outbreak. The other system is MAX software which was used for qualitative data analysis during Kenya's national cholera outbreak of 2015 to describe challenges affecting cholera preparedness of 44 health facilities (Curran *et al*., 2018). Cholera mathematical models which assist in the prediction and analysis of diseases have also been reviewed. One of such models is by Trærup *et al*. (2011) which has integrated historical data of temperature and rainfall with the burden of disease from cholera in Tanzania and also, used social economic data to control the impacts of general development on the risk of cholera (Trærup, Ortiz & Markandya, 2011a). Table 1 summarizes features found in each of reviewed systems and other related papers in this study.

Table 1: Summary of Features Found in Each System Reviewed in this Study

Features	Study 1 (Grad *et al*., 2012)	Study 2 (Pasetto *et al*., 2017)	Study 3) (Koepke *et al*., 2016)	Study 4 (Matsuda *et al*., 2008)	Study 5 (Pascual *et al*., 2008)	Study 6 (Mukhopadhyay, 2015)	Study 7 (Curran *et al*., 2018)	Study 8 (Trærup *et al*., 2011)
Study area	Zimbabwe	Italy	United States	Bangladesh	United States	India	Kenya	Tanzania
Patient full name	√	×	×	√	×	√	√	√
Patient historical details	√	×	√	×	×	√	√	√
GIS-based geographical Home Address	×	√	×	×	×	√	√	√
Date on set	√	√	√	√	√	√	√	√
Time on set	×	×	×	×	×	×	×	×
Rainfall	×	√	×	×	√	×	×	√
Temperature	×	×	×	×	√	×	×	√
Humidity	×	×	×	√	×	×	×	×
Wind	×	×	×	√	×	×	×	×

In general, the studies presented in sections 2.2.1 and 2.2.2, and according to the WHO monitoring and evaluation report (Security & Cluster, 2016) show an improvement of initiatives for ending cholera epidemics in terms of creation of WASH programs and robust community engagements. In addition, the studies show that, most countries affected by cholera especially least developing countries including Tanzania still have (a) fragmented health and environmental systems, (b) lack of uniform data standards, (c) lack of effective tools for timely data analysis and early warnings, (d) no capturing of climatology data in health systems, (e) no quality health-care data, (f) no proper information flow to address the local levels such as district and ward health-care offices, (g) poor and inadequate resources and infrastructures for healthcare activities, (h) lack of clear understanding of the purpose of healthcare data collection, analysis, and prediction and lastly (i) lack of effective tool for prediction of cholera and weather changes linkages (Igira *et al*., 2007; Karimuribo *et al*., 2011; Kwesigabo *et al*., 2012; Kuhn *et al*., 2005). This fact is clearly demonstrated in Fig. 6 by seasonal occurrence and recurrence of cholera epidemics with same causes and transmission patterns in Tanzania (Trærup *et al*., 2010).

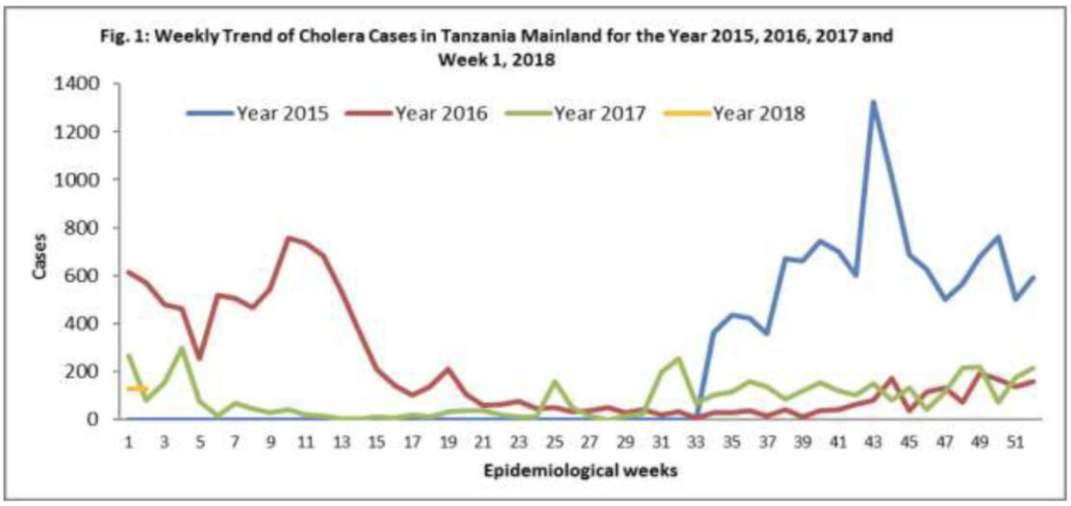


Figure 5: On-going Cholera Epidemics in Tanzania, 2015-2016 by Narra *et al*. (2017)

2.3 Overview of Applications of Traditional Statistical Techniques in Cholera Epidemics

This sub-section identifies and reviews some of the cholera traditional statistics and complex system models. It presents both models which have and have not incorporated or integrated environmental factors in their formulation and then, discusses briefly the strengths and weaknesses in their prediction of cholera outbreaks.

Capasso and Serio are among the very first modellers to formulate the development of a well-known cholera model which has not integrated the aspect of environment as determinants of the prediction and transmission of its outbreak. However, the models assisted not only in the formulation and extensions of other successful and accurate models but also, in provision of knowledge on how the susceptible, infected and recovered population are linked in the transmission of cholera disease (Capasso & Serio, 1978). Several other models have also taken advantage of these models by incorporating additional variables such as education, immunity to the susceptible population, age of the population, and economic factor, to mention a few (Lund *et al.*, 2015). In fact, these very first models which have not incorporated the environmental factors have assisted in the clear understanding of the cholera epidemics in various dimensions. However, they lack the reality touch for not incorporating environmental factors such as water bodies, seasonal weather changes and *V. cholerae* behaviour at certain areas of the environment; which are essential factors in determining the transmission rate, prediction value and causes of cholera epidemic in the area (Jutla *et al*., 2013).

One of the first cholera dynamics models which linked environmental aspects in its formulation was conducted by Codeço (2001). Herein, the model will be referred to as the Codeco model or Threespp model. The Codeco Model is an epidemic mathematical model which considers aquatic reservoir as the only environmental determinants assisting the transmission of cholera epidemics. Table 2 and Fig. 7 illustrate the variables and parameters used in the formula and the compartmental diagram of the Codeco model respectively.

Table 2: Symbols used in the Codeco Model

Symbols	Description
Variables	
S	Number of Susceptible People.
I	Number of Infected People.
B	The concentration of toxigenic *V. Cholerae* in water (cells/ml).
H	Total human population.
N	Human birth and death rates (day-1).
K	The concentration of *V. Cholerae* in water that yields 50% chance of catching Cholera (cells/ml).
Parameters	
A	The rate of exposure to contaminated water (day-1).
R	The rate at which people recover from Cholera (day-1).
Nb	The growth rate of *V. Cholerae* in the aquatic environment (day-1)
Mb	The loss rate of *V. Cholerae* in the aquatic environment (day-1).
E	The contribution of each infected person to the population of *V. Cholerae* in the aquatic environment (cell/ml day-1person-1).

(Codeço, 2001)

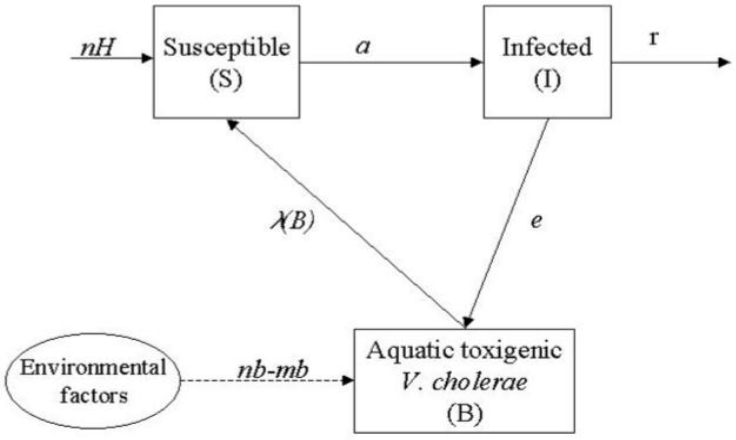


Figure 6: Compartmental Diagram of Codeco Model (Codeço, 2001)

The Codeco model has incorporated very few variables and uses basic Susceptible-Infected-Recovered (SIR) approach coupled with an aquatic population of *V. cholerae*. The objective of this model was to explore the role of the aquatic reservoir on the persistence of endemic cholera as well as to define minimum conditions for the development of epidemic and endemic cholera. Due to the incorporation of the aquatic reservoir as one of the environmental factor or determinants, the model properly explained cholera transmission rate in the community as a product of social and environmental factors. The model has shown the importance of understanding the effect of the aquatic reservoir in the transmission and dynamics of cholera which depends on the sanitary conditions of the community. This study also recommended further extension and development of cholera models in order to understand better *V. cholerae* ecology and epidemiology of cholera. However, given the inability of the model to integrate all essential environmental factors like temperature, rainfall, and humidity levels of the area, it is very difficult to provide a quantitative prediction on cholera dynamics (Codeço, 2001). Nevertheless, the model brings some new insights into cholera epidemiology on essentially quantifying prediction of cholera dynamics into a large aspect or features.

Similar scenarios occur in several other developed and implemented cholera models. Such as; cholera models with a control strategy developed by Wang *et al.* (2011) and Andam *et al.* (2015). These models have simply extended the Codeco model by introducing the variable C as the control factor with the aim of gaining useful guidelines to the effective prevention and intervention strategies against cholera epidemics. The variable C has been used as a control measure to represent the effect of vaccination, therapeutic treatment, and water treatment. The models have also not included vital factors in cholera transmission such as level of temperature, rainfall, humidity, wind, social-economic factors of the patient, geographical location of the area and their surroundings.

Another Cholera epidemic model was formulated by Stephenson (2009) in Zimbabwe. He developed a system dynamics simulation model of cholera outbreaks. The model is, in fact, an SIR model extended with loss of immunity (a return flow from recovered to susceptible after an average of 6 to 10 years), with different degrees of illness, mortality and with a reservoir to harbor the *V. Cholerae*. The formulation of this model contains consequently a number of implicit assumptions such as constant weather seasons throughout the years.

It can therefore be observed that most cholera epidemic models have used traditional statistics and complex system model techniques. These traditional statistics and complex system models are essentially formalization of relationships between variables in the form of mathematical equations (Ledder, 2013). They were invented after mankind realized the limitations of human-brains to store and process large amounts of data (Kelly, 2015). Therefore, these techniques can formulate models that can approximate reality using data reduction techniques. Furthermore, they can describe the different aspects of the real world, their interaction and dynamics through mathematics (Quarteroni, 2009).

Given their capacity, traditional statistics and complex system models have been used in various activities especially in understanding the dynamics of different epidemics. They are however, not without limitations and complexities. To start with, their techniques are based on getting the formulation of a frontier in a classification problem, thus involving a lot of equations and assumptions which sometimes lead to the problem being unrealistic (Richardson, 1979). Secondly, some models are very complex in terms of usability since they need users to have knowledge of mathematical models in order to interpret and understand their result formulation (Cain, 2017). Thirdly, most existing cholera traditional statistics and complex system models have limited the number of variables or parameter in order to be comprehensible and solvable (avoid complexities), thereby decreasing their prediction value. Fourthly, they lack flexibility which is one of the special aspects of any model (Schroeder, Meyer & Taylor, 2013). In addition, the higher the number of datasets or variables, the higher the complexity of the equation in terms of solving, iteration, interpretability, cost as well as computational performance in terms of speed, to mention a few (Foundation, 2007). Lastly, given the current world trend and direction, there is a lot of unmanageable volume and complexity of big data especially in the health sector. Consequently, it is complex to work with cholera traditional statistics and complex system models in some areas due to their limitations (Leskovec, Rajaraman & Ullman, 2014). Therefore, it is high time to explore other techniques in order to support and complement the useful task done so far by the traditional statistics and complex system models.

2.4 Overview of How ML is Being Used in Healthcare Sectors

It has been argued that the adoption of such advanced technologies as ML is rather necessary for the success of the health sector especially in the least developing countries (Kuhn *et al*., 2004). This fact has been proven in several functionalities that ML can perform on data. For

instance, ML can automatically learn and improve from experience without being explicitly programmed (Chao, 2011). Its learning process begins with observing data, such as instructions and direct experiences, in order to observe patterns in the data and make a better decision for future applications. Machine learning evolved from the study of pattern recognition and computational learning theory in artificial intelligence and it is one of the fastest growing areas of computer science, with far-reaching applications, as it teaches computers to do what comes naturally to humans and animals.

Not only that, ML is also closely related to computational statistics and mathematical optimization which also focuses on prediction processes through the use of computers and delivers methods, theories and application domains to the applied field (Du & Swamy, 2014). Furthermore, ML explores the study through construction of algorithms which use computational methods to learn information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Machine learning algorithms are thus used every day to make critical decisions in medical diagnosis, stock trading, energy load forecasting and many other applications (Dietterich, 2009). Moreover, ML's results, prediction and accuracy can be improved through the use of several strategies. Some of these strategies include algorithm tuning, ensemble and extreme feature engineering (Abuassba *et al*., 2017). Furthermore, ML can be applied through the use of python and R software (Lent, 2013). Both of these software(s) are freely available and the codes used are easy to understand and also, they can be easily downloaded to be used and reused based on the case and datasets you wish to solve (Max *et al*., 2017). The following is the review of ML models in terms of its analysis, prediction and imbalance-class.

2.4.1 Machine Learning for Timely Disease Analysis and Prediction

The most effective way of managing disease outbreaks or epidemics is to detect them at early stages so as to control and treat them through proper planning strategies and interventions before they occur or spread further. For this reason, there have been a significant number of studies of ML applications for disease analysis and prediction. This is so, because can quickly formulate sophisticated algorithms to 'learn' features from a large volume of healthcare features and then use the obtained insights to assist clinical practice such as inform proper patient care, reduce diagnostic and therapeutic errors and assist in making real-time

inferences for health risk alert and health outcome prediction (Kelly, 2015). The following is a brief review on how ML has been used in disease analysis

In the review, the study found only three ML models for prediction of cholera, which were done in Iran, Yemen and Uganda. The models for Uganda and Iran will be discussed in this paragraph while that of Yemen will be discussed in Subsection 2.3.2. (Mubangizi *et al*., 2009) in Uganda described methodologies to collect data from different sources and apply ML techniques to predict the risk of cholera outbreak over time in different areas. The study was supported by the IBM Faculty Award. In the study, the data was collected across 80 regions from the Ministry of Health of Uganda. Whereby, the Ministry had a special section in the Health Management Information System (HMIS) which manages Integrated Disease Surveillance and Reporting unit for cholera epidemics. The study formulated a dynamical probabilistic model which can be used to make prediction of cholera infection rates given historical of infection in that area and climatic data (temperature, rainfall and humidity). The other model is by Pezeshki, Tafazzoli-Shadpour, Nejadgholi, Mansourian and Rahbar (2016) in Chabahar city in Iran, which predicted cholera using artificial neural network. The model was capable of forecasting cholera events among 465 villages of the test group with up to 80% accuracy using neural network algorithms. Furthermore, the model managed to show that artificial neural networks are capable of assisting disease forecasts for adopting protective measures using hygienic, social and demographic parameters.

Sarwar (2012) in India, developed Naive Bayes model to predict Type-2 diabetes. In the process a dataset of 415 cases were prepared by collecting the data randomly from different sections of the society. In this study the efficiency of Naive Bayes offered 95% prediction accuracy. The authors opted for a self-data collection in order to have quality datasets. Also, they used only Naive Bayes because from the conducted review in the study it was confirmed in Kononenko's comprehensive comparison that the algorithm (Naive Bayes) outperformed all other 6 algorithms in 5 out of 8 problems of the medical diagnostic domain. Another study that only used Naïve Bayes algorithm is Hoens and Chawla (2013), which proved to analyse heart disease at 86.419% of accuracy using 500 patients' data at diabetic research institute in Chennai. Other studies done in India include; the study that was done by Ramalingam *et al.* (2018) who surveyed five ML models for prediction of heart related diseases or cardiovascular diseases. In this study, they presented a survey of SVM, KNN, Naïve Bayes, DT, RF and Ensemble models. Based on their review, it was concluded that there is a huge

scope of ML algorithms in predicting cardiovascular diseases. In addition the study noted that application of PCA and ensemble multiple algorithms increase the accuracy of the model (Jamgade & Zade, 2019) in India, develop k-mean machine learning model to assist medical officers and systems to early predict patient diagnosis and treatment options. Whereas, (Vasumathi & Kamarasan, 2019) in India, used the DT model to predict the same. The authors proved that ML algorithms produce effective prediction of various disease occurrences in disease-frequent societies. Also, S. Gupta and colleagues (Gupta, Bharti & Kumar, 2019) surveyed ML algorithms like DT, SVM, MLP, Naïve Bayes, KNN and Ensemble classifiers to show that they can lead to rapid disease prediction. Furthermore, (Kadu & Buchade, 2019) proposed the use of KNN and Hidden Markov Model (HMM) to predict non-communicable diseases; and to use fuzzy classification in enhancing the performance of the algorithm. Lastly, Dahiwade and colleagues (Dahiwade, Patle & Meshram, 2019) in India designed a disease prediction model using ML approach using KNN and Convolution Neural Network (CNN). Moreover, Rajeswari and Reena (Sharma & Verma, 2017) diagnose liver disease using FT Tree, Naive Bayes and K star algorithms. Then, the performances of the three algorithms were compared and FT Tree algorithm was selected as the best algorithm among the others in terms of accuracy, performance and speed. The other studies in the next paragraph were done in United States, Greece, Canada and Saudi Arabia to mention a few, which include studies by Frandsen (2016) noted that millions in the United States (USA) suffer from undiagnosed or late-diagnosed chronic diseases such as type-2 diabetes and chronic kidney. Therefore, he used USA's Electronic Record Data to develop type-2 diabetic predictive model using logistic regression, random forest and neural networks in order to catch the disease occurrence earlier and facilitate preventive healthcare intervention which in turn can lead to cost saving and improved health outcomes. Also, Konerman *et al*. (2019) in the USA used national Veterans Health Administration (VHA) data ranging from 2000-2016 to analyse and predict hepatitis C virus. Alotaibi (2019) in Saudi Arabia discussed and implemented a ML models which are; Decision Tree, Logistic Regression, Random Forest, Naïve Bayes and SVM to predict chances of heart failure disease in Saudi Arabia. In his study, he had a limited number of patient's records which were augmented using appropriate techniques. In addition, he was able to increase the accuracy of the models by amplifying the size of the dataset and through the use of a 10-fold cross validation process. (Kourou *et al*., 2015) in Greece present a review of recent ML applications which have been employed to predict cancer progression. In their study, they

stress on the fact that there is a need for an appropriate level of validation in order for these models to be considered in everyday clinical practice.

Aljanabi *et al.* (2018) in Canada reviewed ML classification techniques that have been proposed to help healthcare professionals in diagnosing heart disease. In their review, they found out that most researchers have used data from the same source (UCI repository). In addition, they were able to list down what is required to be done in order to get effective model for accurate prediction of heart disease which include; a dataset with sufficient samples and correct date must be used; the dataset must be pre-processed accordingly because it is the most critical part for getting good results; also, suitable algorithm must be used when developing a prediction model; and lastly it was noticed that ANN performed well in most models for predicting heart disease as well as DT.

Other ML disease prediction studies as summarized in Table 3 which include; the study done on identifying disease- treatment relations using ML approach by (Sairam & Rama, 2016), ML techniques to predict outbreaks and public opinion on health topics from social media by Signorini (2014), intelligent heart disease prediction system using ML by Dessai, (2013) and disease prediction using ML over big data by Krishnan and Kumar (2018), sentimental analysis using ML techniques to predict outbreaks and epidemics by Adhikari *et al*. (2018), malaria outbreak prediction using ML (Sharma *et al.*, 2015; Sharma *et al*., 1988; Wang *et al*., 2019; Kalipe, Gautham & Behera, 2018), deep learning for epidemiological predictions (Chae, Kwon & Lee, 2018) and analysis of prediction accuracy of heart diseases using supervised ML techniques for developing clinical decision support systems (Kiruthikaa & Yuvaraj, 2018) to mention a few.

Table 3: Summary of Reviewed ML Predictive Models

Functionalities/Paper	Challenge	Disease	Weather variable	ML Model	Country (study area)
Signorini (2014)	Fail to provide lead time for optimal public health response	INFLUENZA	×	SVM	USA
Sharma *et al*. (2015)	Lack of early-warning tool	MALARIA	√	SVM and ANN	INDIA
Krishnan and Kumar (2018)	Big Data Analytical Review	HEART	×	CNN	INDIA
Kalipe *et al*. (2018)	Prediction of Malaria using climatic factor	MALARIA	×	KNN, NB and Extreeme GB	INDIA
Kiruthikaa *et al*. (2018)	Analysis of Prediction Accuracy	HEART	√	SVM, DT and NB	INDIA
Adhikari *et al*. (2018)	Epidemic Search Model with Social Network Data Analysis	ZIKA and CHOLERA	×	SVM and NB	INDIA
Chae *et al*. (2018)	Prediction on Big Data	CHICKEN POX	×	DNN	CHINA
Siettos (2018)	Imbalance Class	CHOLERA	√	XGBoosting	YEMEN
Wang *et al*. (2019)	Improve performance of time series models using ensemble method	MALARIA	×	Gradient Boosting Regression Trees (GBRT)	POLAND
Ford and Janies (2020)	Improve performance of ML algorithms using ensemble method	MALARIA	×	DT, Elastic Net, Extreme Random Tree, GB, Lasso Lars, LightGBM, RF and Stochastic Gradient Decent	USA

2.4.2 Imbalance Healthcare Data with ML

In most cases, when working with healthcare datasets, spam filtering, fraud detection and especially data from least developing countries, there is a high possibility of high data imbalance, poor quality, missing information and other uncertainties. Imbalanced classes are a very common problem in ML classification where there are disproportionate rations of observations in each class (Aida, Shamsuddin & Ralescu, 2015). However, the imbalanced data problem can be handled using a confusion matrix which is a table showing correct predictions and type of incorrect predictions, precision, recall, F1-score, sampling and generating synthetic samples techniques (Amin *et al*., 2016). The following a few studies that handled imbalance class; The study done by Aida *et al*. (2015), to formulate a hybrid-techniques by joining generic and Support Vector Machine (SVM) algorithms using wrapper approach in order to analyse five different datasets (Iris, diabetes, breast cancer, heart and hepatitis diseases). The process of joining two algorithms increase the analysis accuracy to 80% of iris disease, 78.26% of diabetes, 76.20% of breast cancer, 84.07% of heart disease and 86.12% of hepatitis disease. The other study was done by (Siettos, 2018) on the on-going Yemen cholera outbreak using a system of four extreme-gradient boosting (XGBoost) ML models which can forecast the number of new cholera cases 2 weeks to 2 months in advance within a margin of just 5 cholera cases per 10 000 people in real-world simulation. To address the need for a complex, reliable model, the author used Feature Engineering and Cholera Artificial Learning Model (CALM); a system of four XGB ML models. However, the model fails to predict sudden spikes due to either lack of information many weeks prior or the events preceding a sharp outbreak not having occurred yet.

Other reviewed studies are summarized in Table 4 which include the study done on ML from imbalanced datasets 101 (Provost, 2000; Hoens & Chawla, 2013) in the USA, presented an overview of sampling strategies as well as classification algorithms developed for countering class imbalance. In this study imbalanced datasets: from sampling to classifiers learning from imbalanced dataset: a comparison of various strategies (Scotia, 2000), experimental perspectives on learning from imbalanced data (Hulse & Khoshgoftaar, 2002), classification with class imbalanced problem (Aida *et al*., 2015), class imbalance problem in data mining (Journal & Science, 2013).

Table 4: Summary of Reviewed ML Models with Imbalance-class

Functionalities/Paper	Data imbalance	Model	Solution	Disease	Country (study area)
Scotia (2000)	√ [Review]	DT, SVM and KNN	Imbalance Class	None	Canada
Provost (2000)	√ [Review]	Naive Bayes, SVM and ANN	Imbalance Class	None	New York
Hulse and Khoshgoftaar (2002)	√	RF	Improve performance of Learners by using 25-Cross Validation and SMOTE	None	USA
Journal and Science (2013)	√ [Review]	SVM,	Feature selection and Sampling using SMOTEBoost	None	India
Hoens and Chawla (2013)	√	Naive Bayes	Wrapper Approach to select sampling strategy and amounts	None	USA
Aida *et al*. (2015)	√ [Review]	SVM and ANN	Data-level Approach [Sampling and Feature Selection] and Algorithm-level Approach [Ensemble Method and one-class learning]	DNA Microarray and Protein Sequence	Malaysia

2.4.3 Discussion and Conclusion for Section 2.4.1 and Section 2.4.2

Lastly, number of studies in the literature review suggested solution to the aforementioned challenges like; having a unique and robust community engagement training, advocacy and communication to all people especially in areas affected frequently with cholera outbreaks; use of mathematical models for prediction and some ML model suggested use of cholera dataset for time series prediction. However, most of existing solutions especially ML models have used datasets which are of good quality, lack missing-information, inconsistency and non-uniform-data-standards, had enough fund to control and manage the data collection in the health sectors and also, had long-time (more than 5 years) measures (plan) were taken with the help of the government in terms of data collection for the specific study (Siettos, 2018; Mubangizi *et al*., 2009; Aida *et al*., 2015; Walldorf *et al.*, 2017; Lugomela *et al.*, 2015; Torres, 2001). In addition, in Tanzania cholera epidemics are very dynamic and comes with a lot of complexity nearly all mentioned in the review such as imbalance-class, missing information, lack of enough dataset, complex weather changes, poor sanitation and hygiene, massive movement of people. Such a situation requires a high level of optimization and regulation in order to develop a suitable model. Furthermore, cholera epidemics in Tanzania take a long time to finish and hence there is overlap of weather variables. Also, in order to

predict cholera epidemics in least developing countries to study its recurring nature based on its varying predictors, which is also a major challenge.

Moreover, most of the reviewed ML models have focussed on using statistical theorems and Artificial Neural Network (ANN). Statistical classifiers have been widely used in prediction diseases such as cardiovascular, cancer and diabetes in the USA, Europe and other developed countries. Based on the nature of the cholera dataset, it will be difficult to work with statistical classifier since it makes very strong assumption on the shape of the data hence most of its prediction is not close to the real situation, in data scarcity, it produces loss of accuracy with prediction equals to zero since if the value of the variable was not seen during model-training data was not seen. In addition, ANN is also difficult to work within least developing countries' settings since it depends on hardware such as parallel processing power in order to operate effectively to mention a few.

Nevertheless, it can be noted that there are quite a limited number of ML applications and the use of ML prediction in cholera epidemics in Tanzania and other least developing countries. This denotes that the use of ML in the health sector especially in the least developing countries is at its infancy stage and this can be seen on the timings of different ML projects, systems, models done in studies such (Mduma, Kalegele & Machuve, 2019; Leo, Luhanga & Michael, 2019). In addition, it was also observed that most hospitals have not implemented the use of ML applications in the Health Systems and also, there is a lot to encounter in the model development process. This situation has been influenced by lack of quality datasets as noted by Pascoe *et al*. (2012) and Kwesigabo *et al*. (2012) and lack of good knowledge on emerging and advanced ML strategies which can work innovatively with any dataset toward fulfilling the intended goal as noted by Raschka (2018), Nejatimoghadam (2018), Domingos (2012) and Chao (2011). This calls the attention of this study to incorporate innovative strategies of technological solutions that can address the existing current challenges of cholera prediction using weather variables in Tanzania. In addition, some of the initiatives and functionalities in these studies, need to be extended to the health sectors in Tanzania and other developing countries in the world where it has been difficult for epidemiological analysts, policy makers and other stakeholders dealing with cholera management to make effective and timely decisions, planning of resource utilization and allocation and also, prioritization of strategies based on location, season, or weather variables (Mghamba *et al*., 2011; Kwesigabo *et al*., 2012; *Tanzania: WHO*, 2015). Therefore, it is high time that the use

of ML techniques is explored in order to facilitate the development of cholera epidemic models in Tanzania's settings.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study Area

Tanzania was selected to be the study area. In the current study, Tanzania was represented by three regions which are; Dar es Salaam, Tanga and Songwe. Tanzania is among the East African developing countries in the SSA which experience frequent re-emergence of Cholera epidemics (Naha *et al*., 2013). For example, between August 2015 to April 2016, Tanzania experienced major cholera epidemics, with more than 14 608 cases and 228 death incidences in the Dar es Salaam region, as shown in Fig. 8 (WHO, 2015). In addition, most cholera outbreaks in Tanzania start in Dar es Salaam region and spread to other regions like Morogoro, Songwe, Tanga and Kigoma (Lugomela *et al*., 2015). However, there are few exceptional cases of cholera outbreaks in Tanzania, which have not started in Dar es Salaam such as; the one that occurred in Kigoma in 2015 as a result of overcrowding crisis of Burundi refugees (Green, 2015). Researchers believed that the dynamics of cholera epidemics in Dar es Salaam region is strongly linked to weather variation (Trærup *et al*., 2010). This is mostly so considering the fact that Dar es Salaam is the most populated region in the country (Tanzania) (Jacobi, Amend & Kiango, 2000). In addition, the region has poor sanitary and hygiene conditions and also, limited resources to sustain people's daily needs. The region therefore, becomes easily vulnerable to the rapid spread of the disease especially in such favourable weather conditions such as during heavy rains (Shemsanga, Nyatichi & Gu, 2010). Furthermore, the country has mostly focused on the use of medical supplies such as water treatment chemicals instead of developing effective models and systems for early prediction and appropriate analysis of cholera epidemics (Trærup *et al*., 2010).

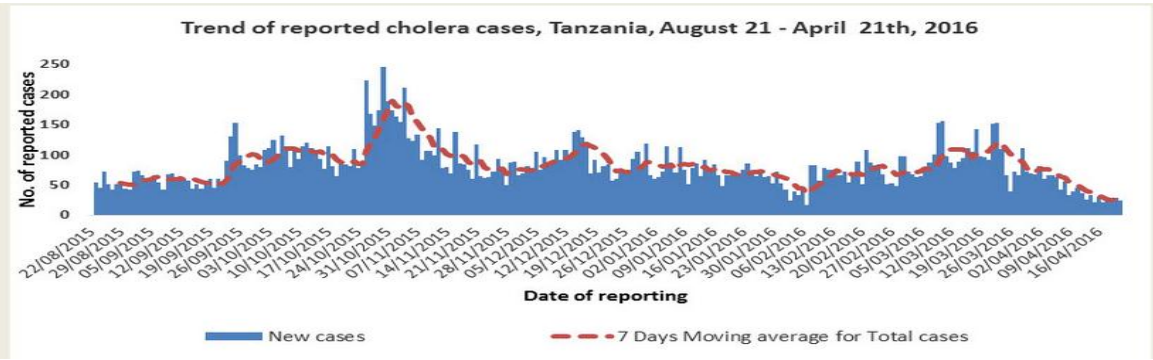


Figure 7: Trend of Reported Cholera Cases in Tanzania, 2016 (WHO, 2016)

Tanga and Songwe regions also have frequent cholera epidemics. Tanga is among the regions in the country with frequent cholera outbreaks similar to Dar es Salaam (Africa, 2018). In addition, most reported cases of cholera which occur in Dar es Salaam always affect Tanga and Morogoro regions. This is so, because of their geographical proximity and also, the three regions (Dar es Salaam, Tanga and Morogoro) have almost similar weather patterns (Yanda *et al*., 2015). Songwe region also gets affected by cholera epidemics. For instance, the May 2018 cholera outbreak in Tanzania affected only Songwe, Arusha and Rukwa regions, out of 26 regions in the Tanzanian mainland (WHO, 2018). Again, in January 2019 for over a period of four weeks, 31 cases of cholera were reported in Kigoma (19 cases) and Songwe (12 cases) (Lwekamwa, 2019). Songwe region is over 800 km away from the Dar-es-Salaam region and it has different climatic conditions (colder weather seasons) to Dar-es-Salaam (Yanda *et al*., 2015). The choice of Tanga and Songwe, apart from the fact that they are also prone to cholera outbreaks, can also help to establish or see on how the developed model performs to regions of a similar (as the case of Tanga) and different (as the case of Songwe) seasonal weather conditions as those in Dar-es-Salaam. Given so, it is apparent that these three regions – Dar es Salaam, Tanga and Songwe as seen in Fig. 9 are actively affected and involved in the transmission of cholera epidemics in Tanzania.

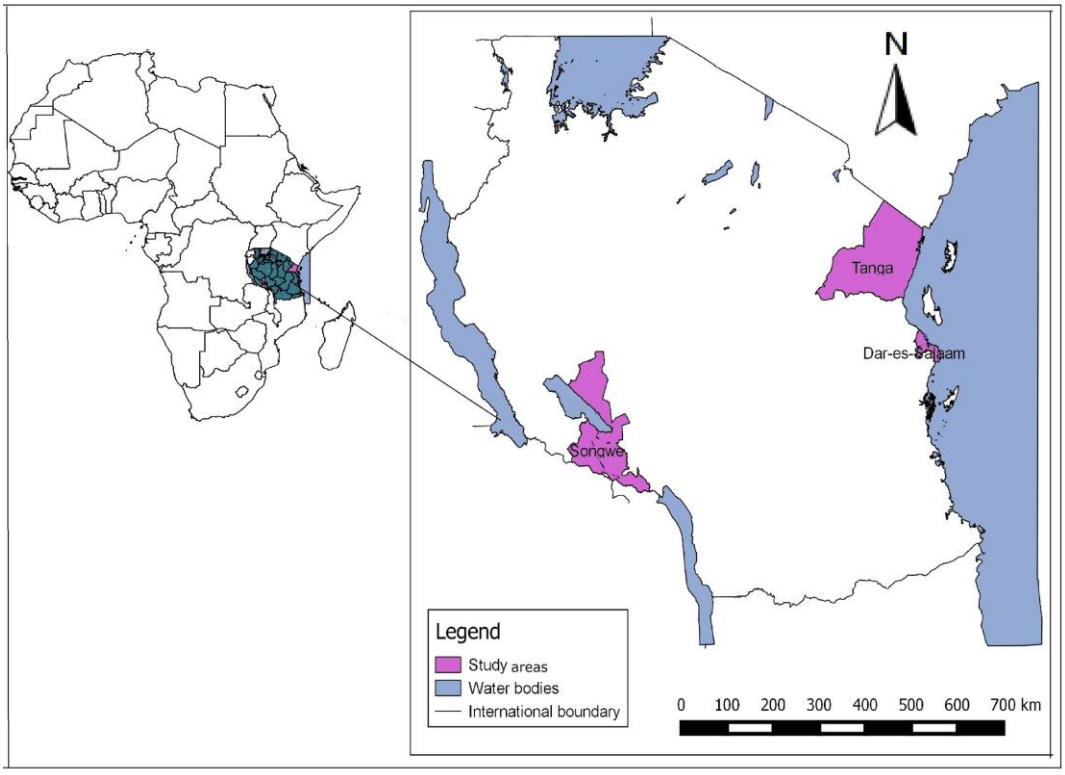


Figure 8: Map of Tanzania Showing Dar es Salaam, Tanga and Songwe Regions

3.2 Development of Climatology-Aware Health Management Information System

During the review of the requirements for the prediction of cholera epidemics, it was observed that most of the health information systems do not collect quality, complete and all essential data. Due to this fact most of the collected dataset for cholera had a lot of missing information and of very poor quality; therefore, it was very necessary for this study to propose as well develop and evaluate a system for linking collection of healthcare and weather variables which will enable future researchers to collect quality data. The system will be termed as Climatology-aware HMIS due to the fact that it collects weather variables which after a long duration of years; the collected weather variables will form climate dataset. The development of Climatology-aware HMIS followed design science research methodology (DSRM) which consists of three stages: (a) problem identification (b) solution design and (c) evaluation and validation. The study chose DSRM because it is outcome-based methodology, which offers guidelines for evaluation and iteration within the project. In addition, DSRM focuses on improving the functional performance of the interfaces (Geerts, 2011).

3.2.1 Problem Identification Stage

In the problem identification stage, we first conducted again a literature review to determine which weather variables are known to be important for Cholera epidemic analysis and prediction, and which of these have been integrated into HMIS in the literature. The study then aimed to determine the status of use of these enhanced HMIS in Tanzanian hospitals. Then, conducted interviews followed by observation session with a total of 30 participants from Dar es Salaam and Kilimanjaro regions. The study selected 30 participants only due to the fact that an intensive literature review on HMIS was done prior this interview session. The participants consisted of 19 males and 11 females, with an average age of 32 years old and an occupation ratio of 60% from the health sector, 30% from the ICT sector and 10% from other sectors. During the interview session, we asked questions such as:

(i) How many environmental factors are included in your HMIS?

(ii) How did you decide on the environmental factors to include into HMIS?

(iii) How do you collect and use environmental data combined with patient diagnosis data?

(iv) What should be done in order to enhance Cholera epidemics analysis linked with environmental factors, particularly the climatology variables?

(v) Is climatology factor essential for Cholera analysis and how?

(vi) What challenges do you face in the currents HMIS towards compiling reports for Cholera epidemics in Tanzania?

Notes were taken during the interview and an inductive coding approach was used during the analysis. Then, the frequencies of responses were tabulated in order to determine the most included features and the reason for their inclusion. We also tabulated responses on reason for exclusion of the other.

3.2.2 Solution Design and Development

Rapid Application Development (RAD) approach was used for solution design because of the short and iterative development cycles with high-quality results of information systems (Lank *et al.*, 2006) that it offers. A total of eight iterations were conducted. Then, the evaluation process of user interface design, features and functionality followed as well as the performance (speed) and security factor using high-fidelity prototypes developed with Java Server Page (JSP), JavaScript and Java codes, MySQL and JSON.

3.2.3 Evaluation of the developed system

For system evaluation, a field study of two months was conducted. At the start of the field study, the following particulars were assessed which include; the aim of system evaluation, the selected study area, suitable months to conduct the study, required number of participants and their role in the hospital through questionnaires. The study consisted of 22 students from Muhimbili University of Health and Allied Sciences (MUHAS) and University of Dar es Salaam (UDSM) who are in the field of ICT and cholera management. The medical students consisted of 10 males and 12 females, with an average age of 30 years old and a specialty ratio of 60% from waterborne epidemics, 30% from the e-Health systems and 10% from medical administration. Questionnaires focusing on whether the developed system will truly enhance the epidemic analysis of cholera with linkages to environmental factors particularly the seasonal weather variables were deployed. Then, data from questionnaires were analyzed statistically using the Statistical Package for Social Science (SPSS) (Arkkelin, 2014).

3.3 Data Collection

In order to achieve the development of a reference ML model for cholera outbreak prediction that incorporates both seasonal weather variables and cholera cases data from the hospital were collected from Dar es Salaam, Tanga and Songwe regions. The weather data for each weather features: temperature, humidity, wind and rainfall, were collected, as the literature shows that these features are the most strongly correlated with cholera outbreaks. Table 5 briefly describes all the collected seasonal weather variables. The data were collected from the Tanzania Meteorological Agency (TMA). Cholera cases data were collected from the Ministry of Health and Social Welfare (MHSW) in Dodoma. Given the sensitive nature of patient data, all patient names and other sensitive patient identities were anonymity. The features covered the area that the patient resided (at district and ward levels), age, gender, cholera-symptoms such as diarrhoea, vomiting and severe diarrhoea, date on set for cholera and patients' laboratory results. Table 5 briefly describes all the cholera cases features included and their format. The date on set variable in this study referred to the date when a patient visited the hospital for *V. cholerae* or cholera diagnosis. This is not only because the incubation period for *V. cholerae* is five days, but also, the range of weather variables within a week is always insignificant (Azman *et al.*, 2013; Werner *et al.*, 2005). In addition, it was assessed by Hoda *et al.* (2011) that, always 60% of cholera patients are at the peak of positive infection (contraction of cholera) on the third day and 72% of cholera patients start showing symptoms on the fifth day. This is the reason as to why the date on set was taken as the day when a patient visited the hospital for the cholera-diagnosis. Therefore, the date on set variable was also collected from the Ministry of Health and Social Welfare in order to assist the alignment of the weather variables to the corresponding patient's details. Additional variables were also included from other sources. These were:

3.3.1 The Patient's Probable Source of Drinking Water was Labelled Water Variable, with the Possible Values being Treated or Clean Drinking Source of Water (1) and Untreated or Dirty Drinking Source of Water (0)

The water variable was collected based on the fact that cholera is a waterborne disease and mostly transmitted through contaminated food or drinking water. The Water Variable lacked self-reports of drinking water sources of nearly 30% therefore, the study hypothesised that clean sources of drinking water can be represented by water whose source is from Dar es Salaam Water and Sewerage Corporation Authority (DAWASA) and the Ministry of Water

as the two are committed to providing affordable, clean, safe water and effective sanitation and sewerage services to its customers. In addition, water from DAWASA and Ministry of water includes people who use registered well water. This is because in Tanzania, there are a lot of people who use water from the well for drinking. However, the water from wells must comply with the standards set by DAWASA and the Ministry of Water for it to be considered as clean water. Patients with missing water sources data, who resided in areas where DAWASA or ministry of health were in operation at the time of contracting disease, were labelled as having clean water.

The study has hypothesised the 30% the water variables values which were missing in the dataset, based on the fact that there are several cholera mathematical models developed which have used aquatic reservoir and other water bodies such as river in the area to represent the status of drinking water and sewerage level (Codeço, 2001) and (Stephenson, 2009). In addition, the study did not use the patient's Salary (Income) variable to support the distribution of water variable because it (Salary variable) is distant from Water Variable as shown in Fig. 10 and Fig. 11 and also, is an outlier as shown in Fig. 12. Outlier is a data point in a dataset that is distant from all other observations or in other words a type of data or variable which cannot smoothly assist the prediction of cholera epidemics. Furthermore, according to the literature, it has been noted that cholera is influenced by people's cultural practices, attitude and behaviour such as wedding, burial and open defecation; meaning that a person can have clean water and high salary but still get or experience cholera disease due to his cultural practice, attitude or behaviour.

3.3.2 Political Economy operating in the Patient's District which was collected from the Ministry of Home Affairs

The study hypothesized that this data (Political Economy) can assist to understand the inter-relationship between political processes, economic systems and health systems of the patient's area. The dimension of the political economy could assist in knowing the attitude and nature of people in the area in terms of supporting government's health sectors initiatives such as; protection of sources of water, cleanliness of surroundings and climate change mitigation like *kata mti, panda mti; meaning "if you cut a tree, then you must plant a tree"* (Peterson, 2008). In addition, other studies show that sometimes political affiliation can affect co-operation with the government initiatives such as sabotage of government initiatives and policies (Narra *et al*., 2017).

Lastly, the seasonal and cholera cases data collected for the different regions covered different time periods. This is so because, these time periods where the time when these regions experienced cholera and also, that was the only available data from the Ministry. Data for Dar-es-Salaam and Tanga regions covered the January 2015 to December 2017 time period, while data for the Songwe region covered January 2017 to December 2018.

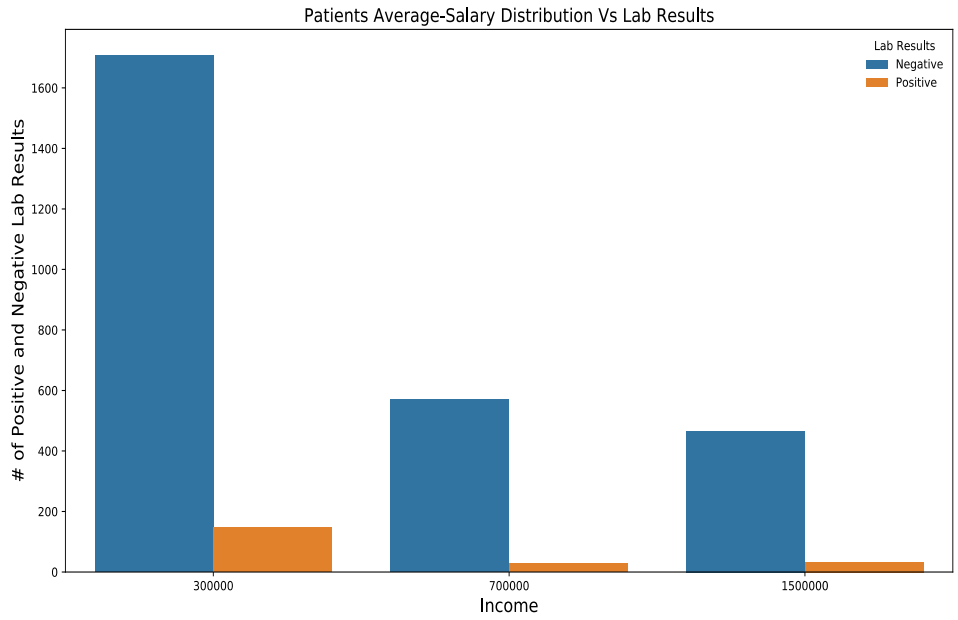


Figure 9: Patient's Average Salary Distribution vs Lab Results

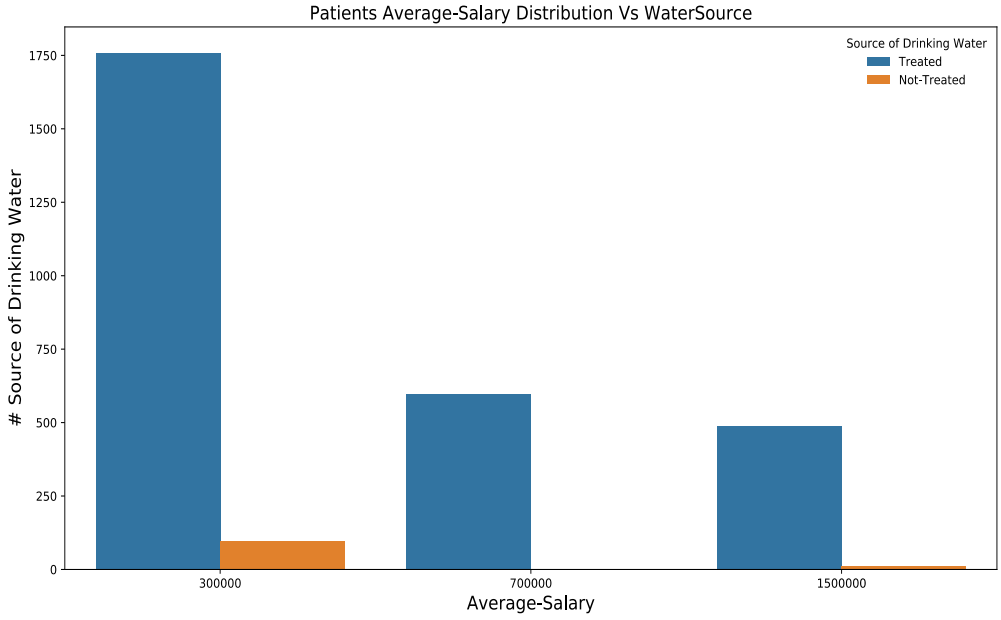


Figure 10: Patient's Average Salary Distribution vs Source of Drinking Water

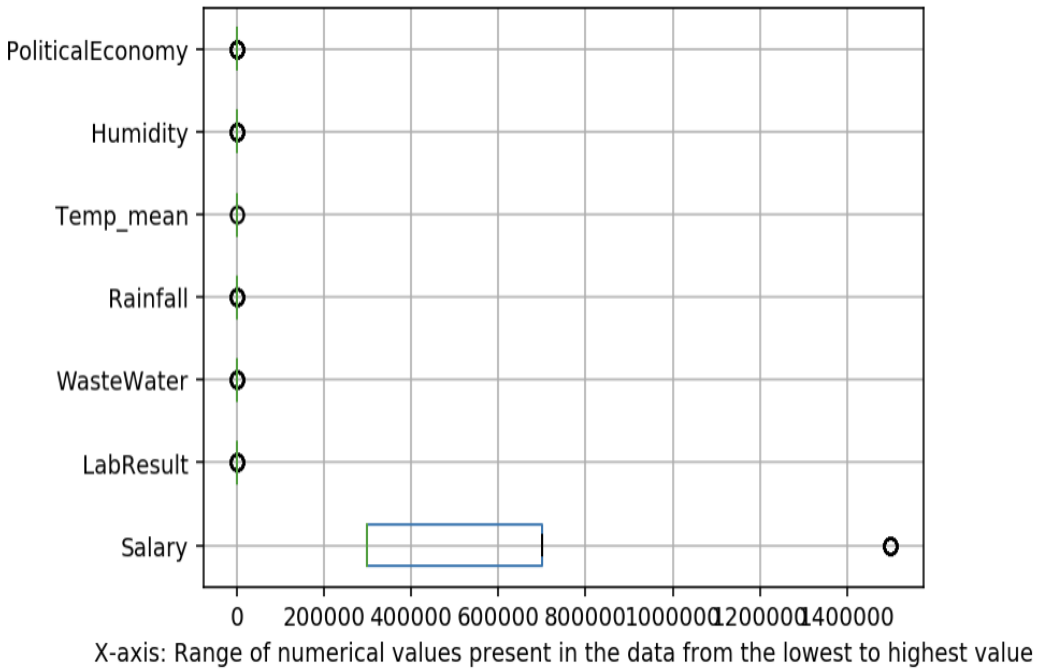


Figure 11: Patient's Average Salary as an Outlier

Table 5: Collected Cholera Dataset Description

Variables	Description	Measurements of Variables
Daily Seasonal Weather Changes Data		
Temp_max	Minimum Temperature	Degree Celsius (°C)
Temp_min	Maximum Temperature	Degree Celsius (°C)
Temp_mean	Mean Temperature	Degree Celsius (°C)
Rainfall	Rainfall	Millimetre (mm)
Humidity	Relative Humidity	(%)
Wind_Spd	Wind Speed	Knots
Wind_Dir	Wind Direction	Degrees
Cholera Cases Data with Regards to Patient Details		
District	District location	Dar es Salaam Districts
Ward	Ward location	Dar es Salaam Wards
Age	Age	Numbers
Sex	Sex	Female (F) or Male (M)
Symptoms	Symptoms	Diarrhoea or Vomiting or Severe Diarrhoea (S. Diarrhoea)
DOS	Date on set	Date-Month-Year
Results	Lab result	Positive or Negative
Outcome	Outcome	Alive or Died
Occupation/ Salary	Occupation/ Salary	Type of Occupation/ Average Salary (Tshs)
Water Source Type		
Water	TreatedWater / CleanWater	1
Variable	UnTreatedWater / DirtyWater	0
Political Economy (PE)		
Positive	People fulfil PE's requirements	1
Negative	People do not fulfil PE's requirements	0

3.4 Data Pre-processing

Data pre-processing process involved several techniques like cleaning, integration, transformation and reduction. The collected dataset from Dar es Salaam, had patients' data of

6800 observations with 18 variables. The 18 variables of the patients included; region, district, ward, age, sex, occupation, data on set, diarrhoea symptoms, vomiting symptoms, severe diarrhoea symptoms, laboratory result, outcome, water sources, eating place, rainfall, temperature, humidity, wind and political economy. Based on the literature and nature of the data the following pre-processing techniques were performed;

The first stage of data pre-processing was data cleaning which recognizes typographical errors, incorrect, imprecise, partial, inappropriate parts of the data from the datasets. The cleaning process also involved ignoring tuples which contained missing values and alter values compared to a known list of variables in the dataset. Given the cleaning process, there was no misspelling error, but there were 32 missing weather data. This was addressed by visiting TMA offices in order to cross-check the received data so as to fill the missing values. Duplicates were identified and cleared and data inconsistencies were corrected in order to obtain similar values or observation of a variable. Precisely, data cleaning consisted of the following steps as shown in Table 6. All steps were run repetitively on the data till all problems related to data quality were solved.

Table 6: Data Cleaning Stages

Steps	Description
Data Analysis	Dirty data detection by reviewing the dataset and quality of data.
Define Workflow	Define the cleaning rules which included; use of expectation maximization in handling missing values.
Execute defined rules	Rendering the defined rules on the source dataset process and display result in clean data.
Verification	Verify the accuracy and efficiency of cleaning rules with the responsible offices, literature and user requirements.

The second stage of data pre-processing was data integration which assisted in merging data derived from different sources of data into a consistent dataset. In the process the integration process followed uniformity of standard measure as described in Table 5. Then, all files were integrated using advanced Microsoft Excel techniques.

The third stage of data pre-processing was data transformation which assisted in transforming all data into a format suitable for analysis. For example, values of LabResult variables which were positive and negative were transformed into 1 and 0 respectively.

The fourth stage of data pre-processing was data reduction. One of the rules that were applied in the data reduction process was to reduce data without loosening the integrity of the dataset. For example, the following data or variables values were reduced; region, age, sex, occupation, diarrhoea symptoms, vomiting symptoms, severe diarrhoea symptoms, outcome and eating place. The following are some of the reasons which guided the reduction process; region was removed since data from the 3 different regions were given different files. The ward variable was removed due to the fact that it had already assisted the alignment of water distribution in the district level and therefore in the place of patient's location district variable was used since the ward variable also had a lot of missing information at a rate of 78%, according to the literature (WHO, 2007) patient's age sex had no significant linkage to cholera-laboratory-results, all cholera patients had vomiting, diarrhoea and severe diarrhoea symptoms, therefore, the symptom variables were represented by the LabResult variable (Meaning, if it is yes cholera then the patients had the mentioned symptoms and vice versa), outcome (dead or alive) had a lot of missing information, eating place had the same answer overall as home (i.e. home) at a rate of 100% hence, it had no effect on the laboratory result (LabResult).

Lastly, the pre-processed dataset was stored in Microsoft Excel spreadsheet of Microsoft Office 2013 suite of desktop publishing (.xls). All the pre-processing stages were performed using Microsoft Excel techniques. Lastly, the cleaned dataset remained with 6218 patients' data (observations) with 8 variables which include; Temperature, Rainfall, Humidity, Wind, District, LabResult, WasteWater and PoliticalEconomy.

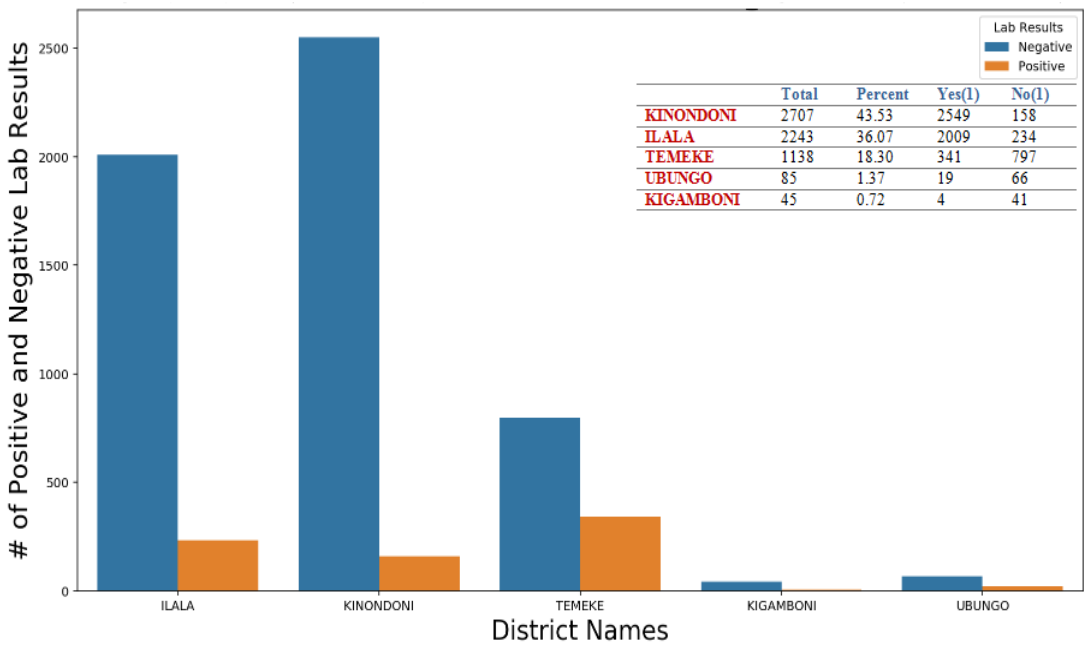
3.5 Statistical Data Description

The dimensions of the dataset were calculated using descriptive analysis in order to determine the normality of the distribution. Therefore, the pre-processed dataset was then described in terms of counts, means, standard deviations (std), minimum (min), maximum (max), 25^{th}, 50^{th} and 75^{th} percentile in order to understand their dimensions, as shown in Table 7. This is so because this is a non-normal distribution and therefore, non-parametric statistics tests are needed. The table count shows a total number of pre-processed-data in each column; mean shows the average value of each column; min and max show the minimum and the maximum number of each column respectively, and std. shows the standard deviation of each column (Leys *et al.*, 2013). Figure 13 to Fig. 18 provide a graphical representation

(histograms and line graph) of the pre-processed data. Figure 13 presents patient's distribution per months; Fig. 14 presents clean and waste water distribution per districts; Fig. 15 shows rainfall distribution per months; Fig. 16 shows patient's distribution across districts, Fig. 17 presents the total rate of clean and waste water distribution and Fig. 18 presents the rate of outlier for the cholera dataset. The outlier is a data point in a dataset that is distant from all other observations or in other words a data point that lies outside the overall distribution of the datasets. In the current study, outliers are the types of data or variables which cannot smoothly assist to get its linkages with the occurrence of cholera epidemics. They do cause some implications such as they can deviate the answer if used to generate a particular relationship or cause errors as they do not have proper match with the resultant data (cholera). Therefore, as described in Fig. 18, the outlier is Wind Direction.

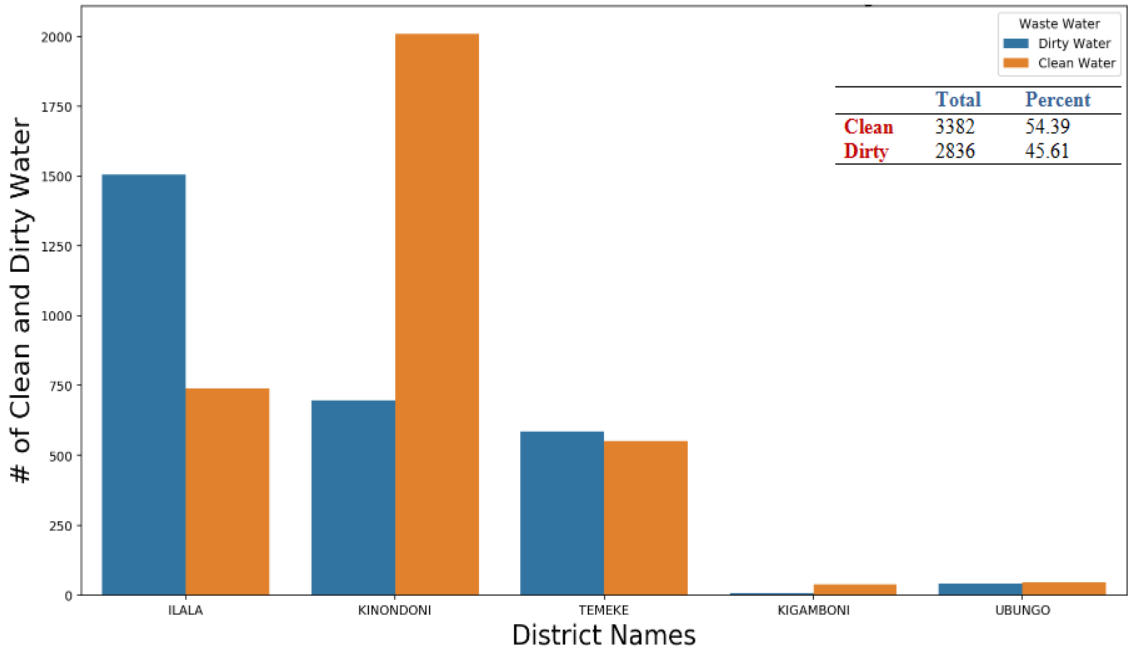
Table 7: Statistical Data Description of Cholera Cases Using Count, Mean, Std, Min, Max and Percentile

	Count	Mean	std	min	25%	50%	75%	max
Rainfall	6218	1.746	6.623	0	0	0	0.1	105.1
Temp_max	6218	31.342	1.688	0	30	31.1	32.7	36.3
Temp_min	6218	22.528	2.261	0	21	21.4	24.1	28.8
Temp_mean	6218	26.935	1.667	0	25.5	26.7	28.2	31.55
Temp_range	6218	8.814	2.194	0	7.7	9	81	16.4
Humidity	6218	78.954	4.898	0	76	78	81	98
Wind_Dir	6218	113.659	79.13	0	60	120	160	360
x_wind	6218	0	0	0	0	0	0	0
y_wind	6218	0	0	0	0	0	0	0
WasteWater	6218	0.544	0.498	0	0	1	1	1
PoliticalEcon	6218	0.714	0.261	0	0.5	0.5	1	1
LabResult	6218	0.122	0.327	0	0	0	0	1



	Total	Percent	Yes(1)	No(1)
KINONDONI	2707	43.53	2549	158
ILALA	2243	36.07	2009	234
TEMEKE	1138	18.30	341	797
UBUNGO	85	1.37	19	66
KIGAMBONI	45	0.72	4	41

Figure 12: Patients Distribution across Districts



	Total	Percent
Clean	3382	54.39
Dirty	2836	45.61

Figure 13: Clean and Waste Water Distribution per Districts

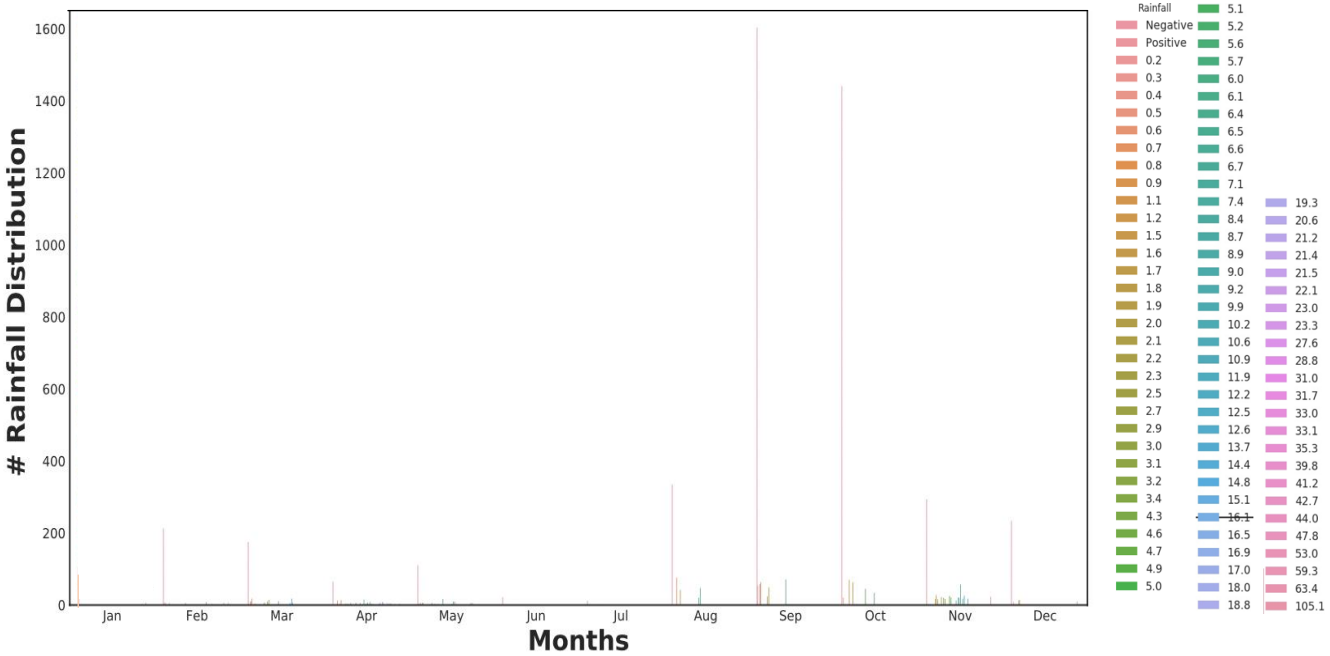


Figure 14: Rainfall Distribution per Months

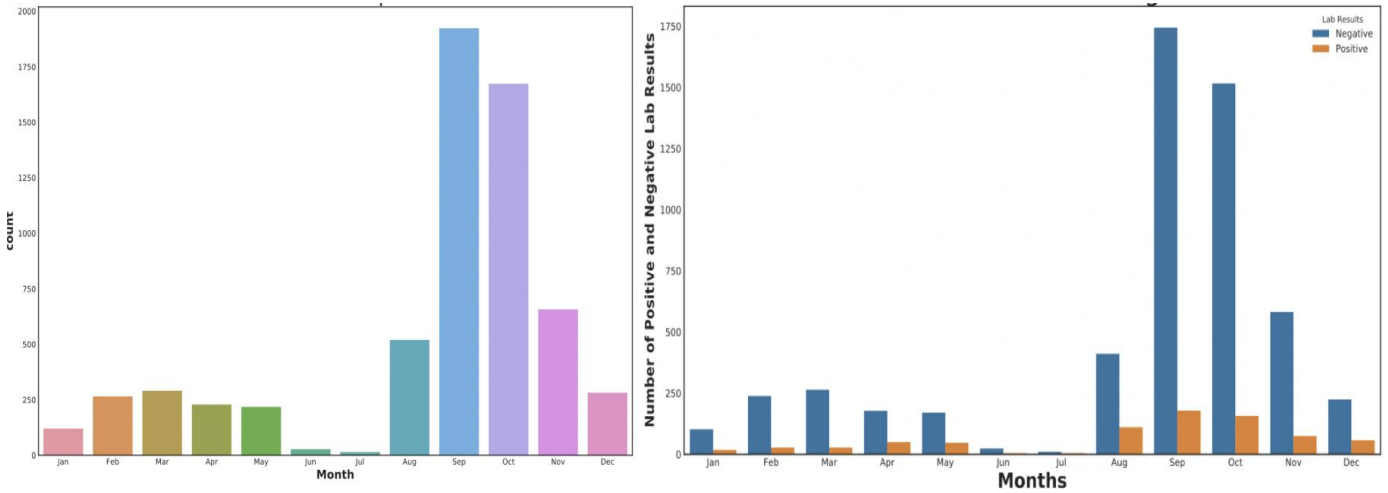


Figure 15: Patient Distribution per Months

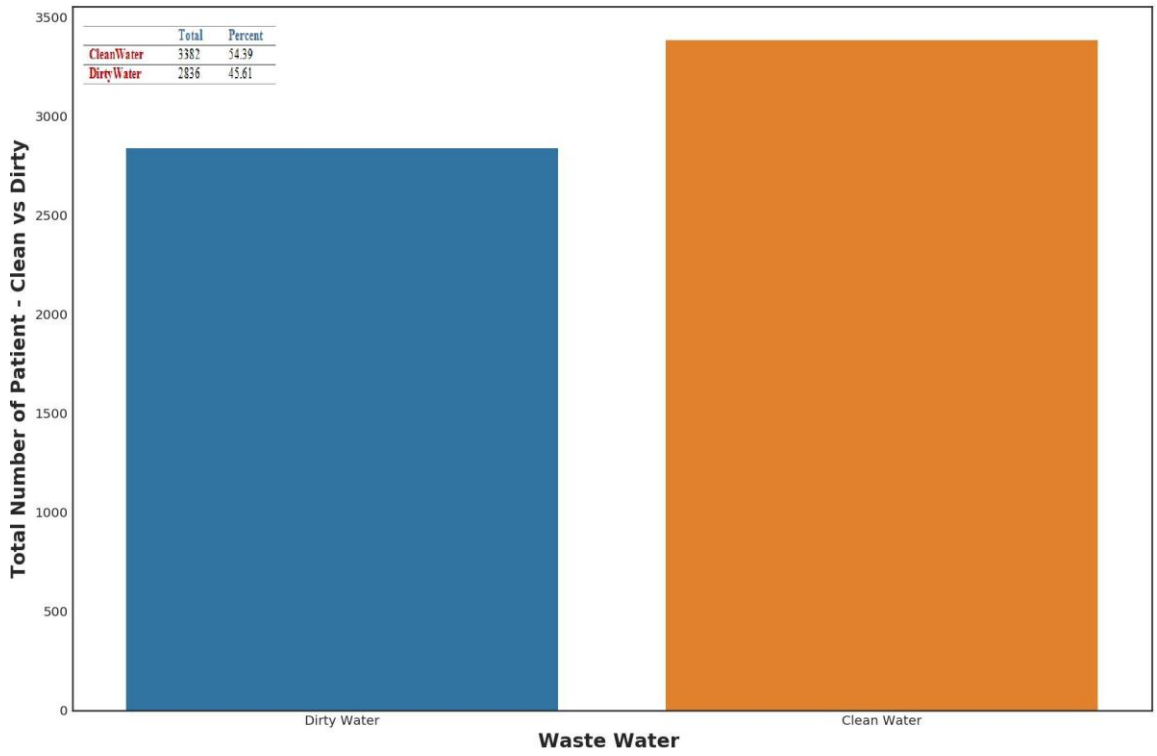


Figure 16: Total Rate of Clean and Dirty Water Distribution

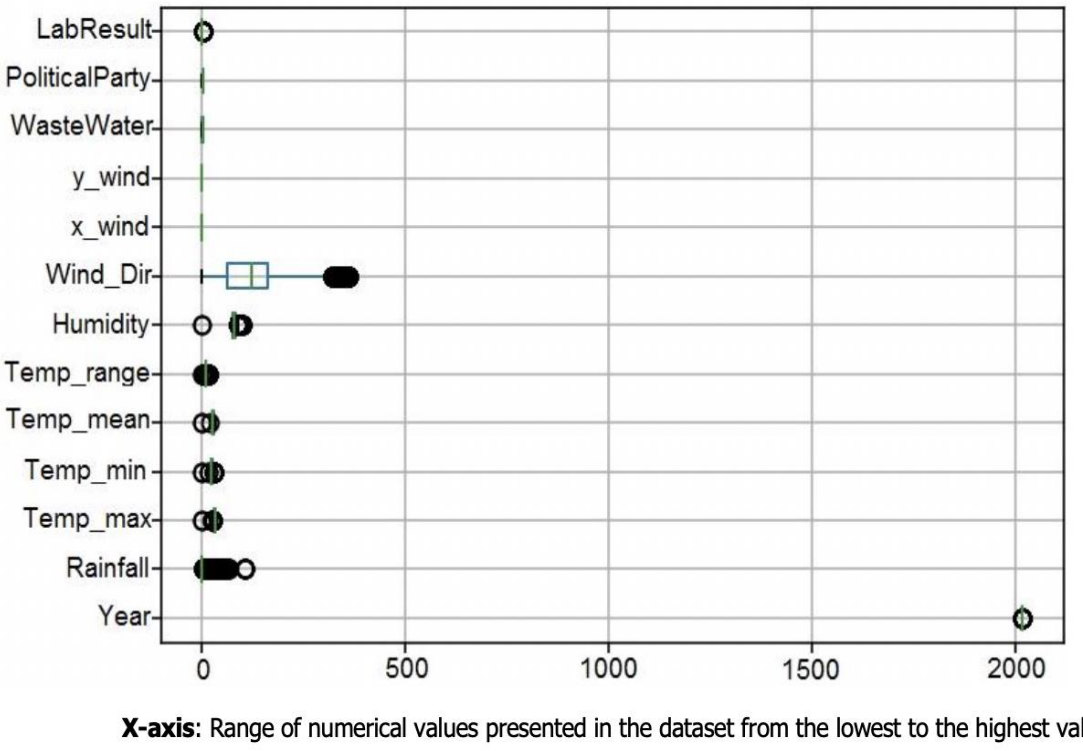


Figure 17: Outlier for Cholera Dataset

3.6 Model Development Approach

In order to achieve the intended goal of developing the proposed model, the procedures briefly illustrated in Fig. 19 were followed. The following paragraph illustrates and explains each step used in the development approach in Fig. 19.

In this study, the final dataset after pre-processing and feature selection was referred to as cholera dataset. The Scikit-learn library for the Python programming language was used for developing the model. The python language was used due to the fact that it is open source with wide community development, easy to learn, has extensive support libraries and contains numerous third-party modules for interacting with other languages and platforms (Geerts, 2011). In the development approach, firstly the balance-level of cholera dataset in terms of its class (no cholera and yes cholera) was tested/ checked and it was found-out that the cholera dataset was imbalanced. The balancing was performed in order to increase the effectiveness of performance of a classifier (Amin *et al*., 2016). Then, a sampling procedure was performed on the data in order to restore the balance. Thereafter, 30-fold-cross validation process was performed on cholera dataset in order to have much iteration and also, reduced variability, overfitting and selection bias (Domingos, 2012b). The 30-fold-cross validation as a test method was performed randomly on chunk of 70% training and 30% testing of cholera dataset. The dataset was split in 70% and 30% as training and test set respectively, in order to avoid model-overfitting and under-fitting (Konerman *et al*., 2019). In the process, the training data was used to build the models while the testing data was used to assess the prediction performance of the models. After building the model using 10 ML algorithms, evaluation metrics were performed in order to select the best performing models or algorithms (Kalipe *et al*., 2018). The process used 10 ML algorithms in order to have a wide range for selection of the best model. In addition, ensembled classifiers such as bagging classifiers were used among the 10 algorithms in order to increase the performance, accuracy, reduce variance and prevent overfitting and underfitting of the final model (Signorini, 2014). This is so because it was observed earlier that RF algorithms were performing well on the dataset. Based on model development approach as displayed in Fig. 19, if more than one model was selected as the best models, model tuning process was performed and then the same process was repeated from model evaluation in order to confirm their authenticity before the process move into model evaluation process; and if only one algorithm was selected then the process also, move into model evaluation process. Overall, the used model development approach is a novel

approach and it was tailor made specifically for the nature of cholera dataset in this study. The following Sections, explain into details on what and why transpired in each step of model development approach.

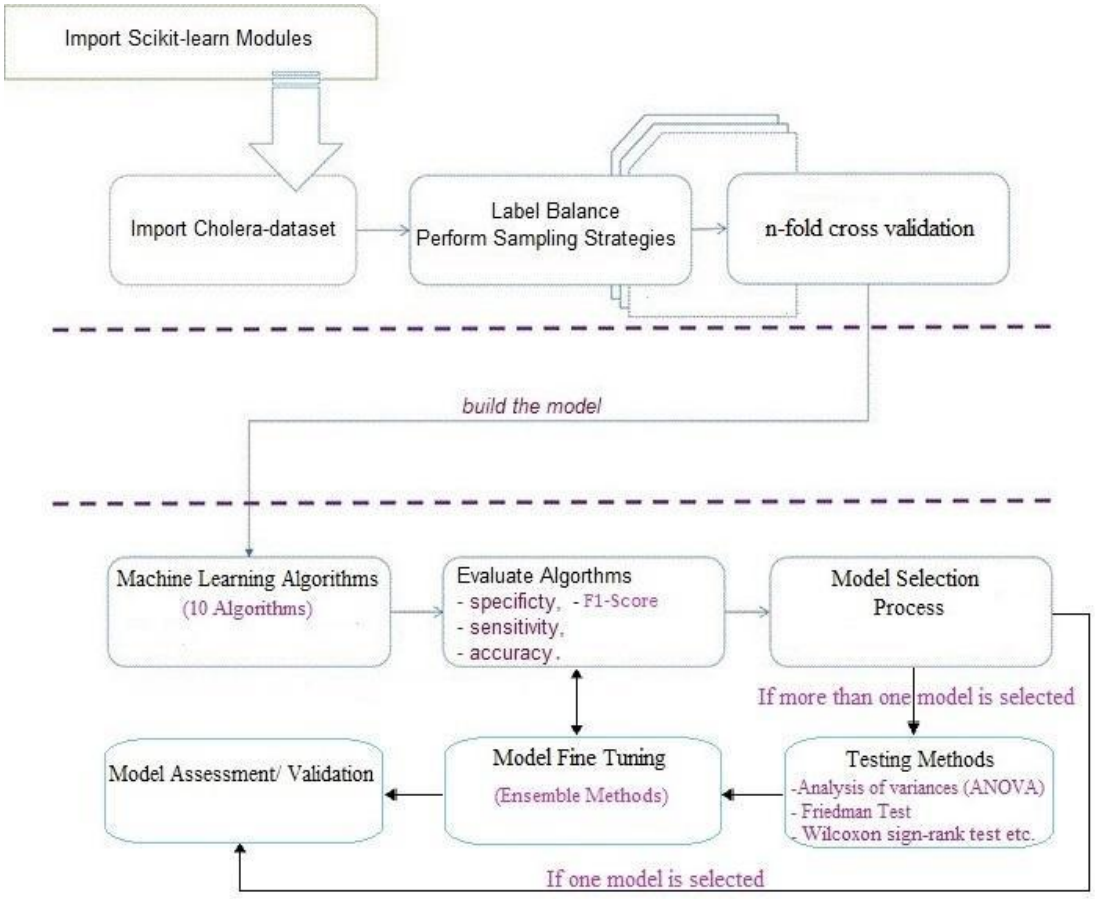


Figure 18: Model Development Approach

3.7 Machine Learning Models

In ML, there is no one algorithm that works best for every problem, as there are many factors at play, for instance, the size and structure of the datasets (Kalipe *et al*., 2018) and (Signorini, 2014). In order to determine which classification algorithms would be best suited for predicting cholera disease outbreaks based on weather variables, ten best classification algorithms for disease prediction (based on literature) were selected for performance comparison. These algorithms were: XGBoost, Gradient Boosting, Random Forest (RF), Bagging, Multi-Layer Perceptron (MLP), K-Nearest Neighbours, Decision Tree (DT),

Support Vector, ExtraTree and Linear Regression (Kotsiantis, 2007). Following hereunder is a brief descriptive discussion of these algorithms.

In a nutshell, in ML there is no one algorithm that works best for every problem, as there are many factors at play, for instance, the size and structure of the datasets. Therefore, ten best supervised ML algorithms for disease prediction were selected to conduct the study. The study considered using all the ten ML algorithms in order to select the best performing algorithm among the wide range. These algorithms include: XGBoost, Gradient Boosting, Random Forest (RF), Bagging, MLP, K-Nearest Neighbours, Decision Tree (DT), Support Vector, ExtraTree and Linear Regression (Kotsiantis, 2007). Following hereunder is a brief descriptive discussion of these algorithms.

Briefly, DT is a classification and regression algorithm which uses tree-like mode or flowchart-like tree structure to generate decisions (Sharma & Agrawal, 2013). A DT typically start with a single node, which branches into possible actions against one another based on their costs, probabilities and benefits. The nodes are categorized into chance nodes which are represented by circles, decision nodes which are represented by squares and end nodes which are represented by triangles. As an example, one can consider the following simple investment decision model. Imagine having $25 which can be invested in two companies, A and B. Companies A and B can give an investment outcome of $125 and $90 respectively. If things go positive or negative in companies A and B, there is a likelihood of getting an income or loss of $100 or -$50 and $200 or -$100 respectively. Now, based on the branches of the DT, one has to decide in line with his/her requirements, as shown in Fig. 20.

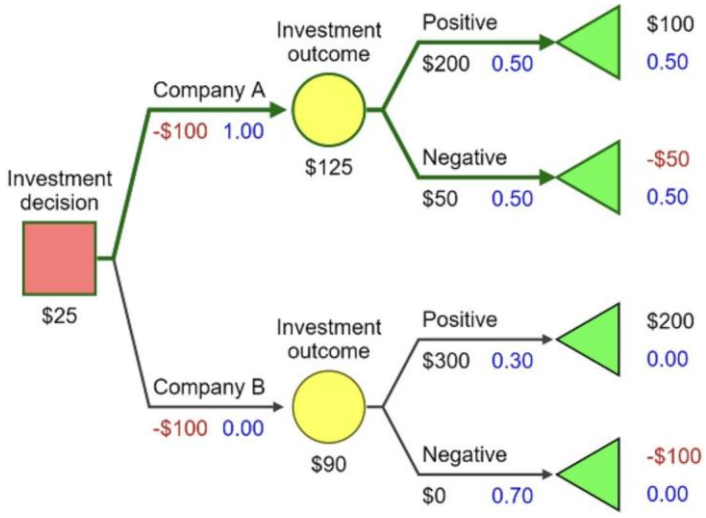


Figure 19: DT Classifier (Sharma & Agrawal, 2013)

Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregates their individual predictions either by voting or by averaging to form a final prediction as shown in Fig. 21. In bagging classifier, voting is always applied for classification and averaging is for regression problems (Abuassba *et al*., 2017).

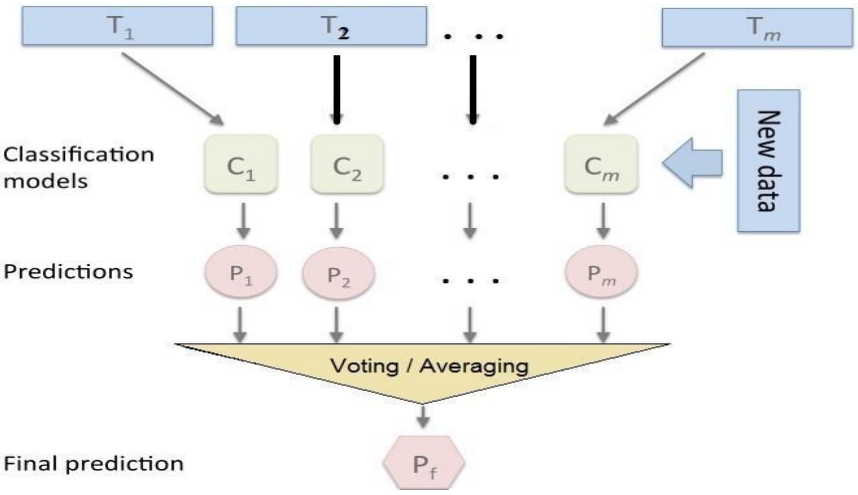


Figure 20: Bagging Classifier (Sutton, 2005)

Random Forest is a meta-estimator which fits a number of decision trees on various sub-samples of the dataset and then averages the results in order to improve the predictive accuracy and control over-fitting (Peng, 2015). Each individual tree in the RF splits out a class prediction and the class with the most votes becomes the model's prediction, as shown in Fig. 22.

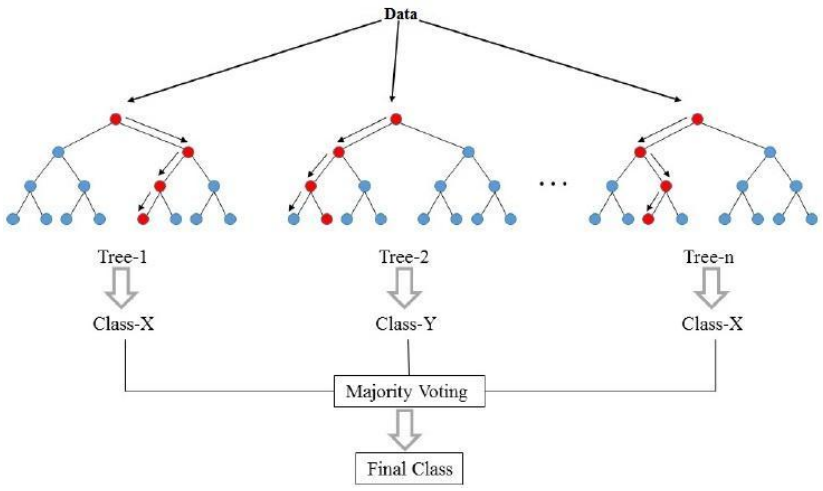


Figure 21: Random Forest Classifier (Louppe, 2014)

ExtraTree classifier is a meta-estimator that fits a number of randomized decision trees on various sub samples of a dataset and uses averaging to improve the prediction accuracy and control over fitting (Louppe, 2014). It is like RF which builds multiple trees and splits nodes using random subsets of features, but with two key differences, namely; it does not bootstrap observations and nodes are split on random splits.

Gradient Boosting is one of the most powerful techniques for building predictive models in the form of an ensemble of weak prediction models, typically decision trees (Beygelzimer *et al.*, 2015; Natekin & Knoll, 2013). Whereas, XGBoost is also a decision-tree-based ensemble ML algorithm which uses a gradient boosting framework, as shown in Fig. 23. It is highly efficient, flexible, portable and provides a parallel tree boosting that solve many data science problems in a fast and accurate way (Max *et al.*, 2017).

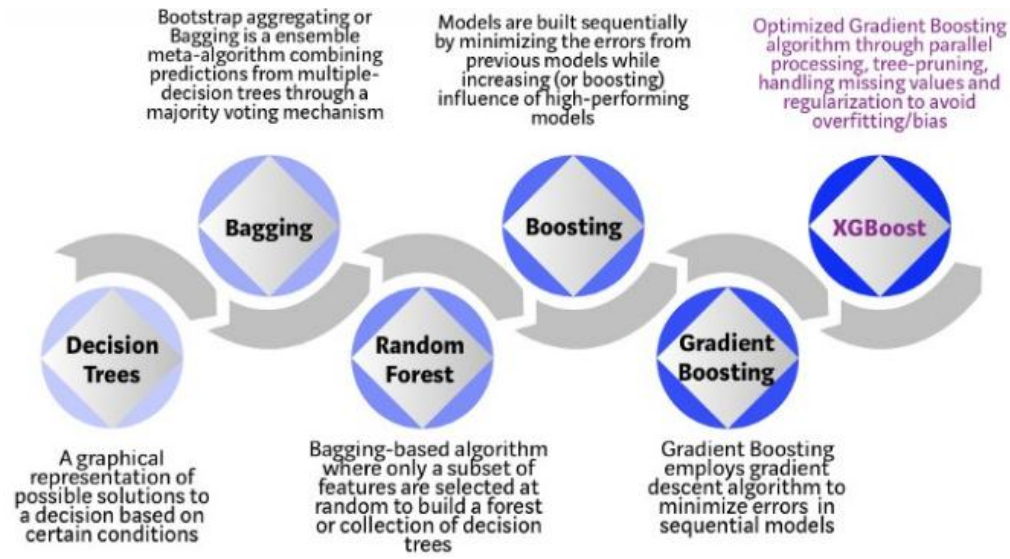


Figure 22: Evolution of XGBoost from DT Classifier (Mitchell & Frank, 2017)

Multi-Layer Perceptron is a deep, artificial neural network which consist of more than one perception. They are composed of an input layer to receive the signal, an output layer that makes a 'decision or prediction' about the input and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP, as shown in Fig. 24. The hidden layer is capable of approximating any continuous function (Wang, 2018). The MLP is usually applied to supervised learning problems in order to train on a set of input-output pairs and learn to model correlation between inputs and outputs (Yeh *et al.*, 2011).

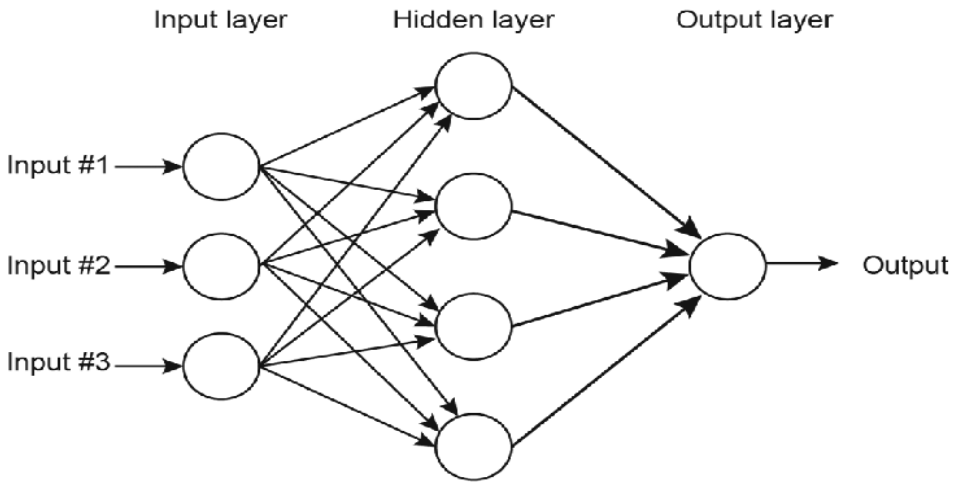


Figure 23: MLP Classifier (Breve, Ponti & Mascarenhas, 2007)

K Nearest Neighbours is a very simple supervised algorithm used in data mining and machine learning. It's a non-parametric classifier algorithm where the learning is based on how similar a data is from another (Beyer *et al.*, 1999). To explain this algorithm, consider this simple case. The following is a spread of red circles (RC), green rectangles (GR) and a blue star (BS). Blue star can either be RC or GR and nothing else. Measure the distance (K) using Euclidian, Manhattan, Minkowski or weighted, say K=3. Then, make a circle with BS at the centre just as big to enclose only 3 data points on the plane. Based on the data points, the three closed points are all RC. Hence, the choice becomes very obvious that BS belongs to RC, as shown in Fig. 25.

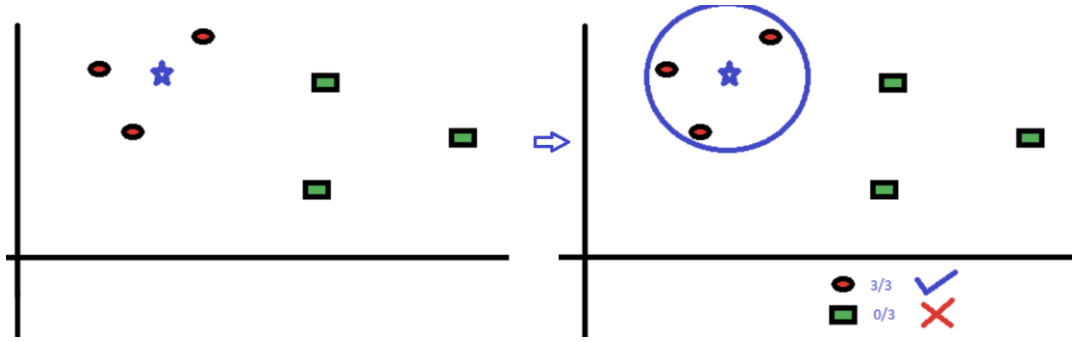


Figure 24: K-NN Classifier (Domingos, 2012a)

Support Vector Machine is a supervised machine learning algorithm which can be used for both classification and regression analysis (Wilcox *et al.*, 2013). For example, take a set of training examples, each marked as belonging to one or the other of two categories. The SVM training algorithm builds a model that assigns new examples to one category or the other,

making it a non-probabilistic binary linear classifier. In a labelled data, SVM algorithm can output an optimal hyper plane which categorizes the data into two dimensional spaces, as shown in Fig. 26.

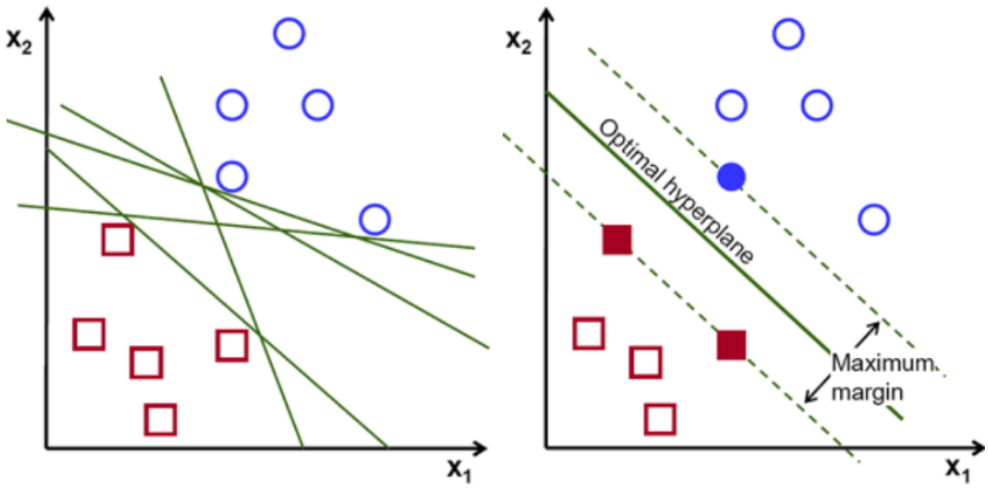


Figure 25: Support Vector Machine Classifier (Zisserman, 2013)

Lastly, Linear Regression (LR) is a supervised machine learning algorithm widely used for data analysis. Linear Regression achieves making classification decision based on the value of a linear combination of the characteristics. In LR, the value of y is given as y = m*x + c; where m is the gradient of the line or slope of the line, x is the input value and c is the intercept value, as shown in Fig. 27 (Huang & Fang, 2013).

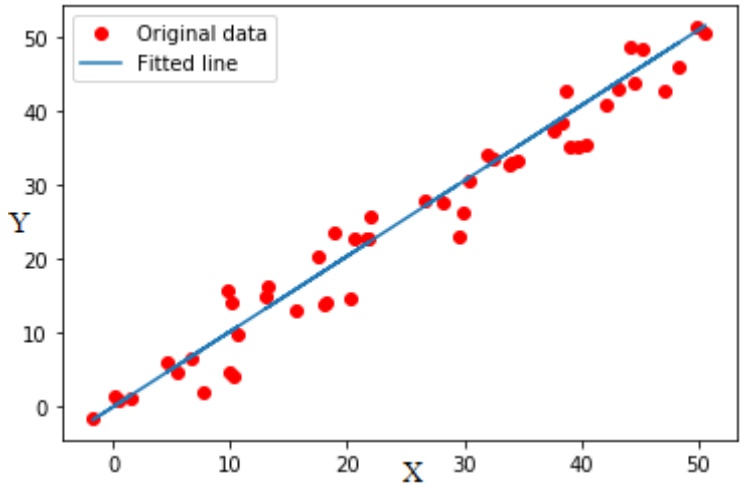


Figure 26: Linear Regression Classifier (Huang & Fang, 2013)

It can be observed that Random Forest, Bagging and ExtraTree classifiers are improvements of other algorithms (Abuassba *et al.*, 2017). The study has therefore used enhanced (ensembled) algorithms due to the fact that prior works have shown that although simple classifiers are widely used, ensemble ones tend to be more accurate. In addition, in this work both simple and ensemble classifiers were included in order to determine whether this applies to the cholera cases as well. Furthermore, the used ensemble classifiers are products of DT which each have been optimised to address challenges of using DT as a predictor. These challenges or drawbacks of DT include high variance, low prediction accuracy and overfitting, to mention a few (Peng, 2015; Louppe, 2014; Biggio *et al.*, 2011; Raschka, 2018; Juliussen, 2009).

3.8 Data Imbalance Problem

The collected dataset was imbalanced at a rate of 0.122 as shown in Fig. 28. The majority class which is **NO Cholera (0)** dominated the prediction value. This means classification of the minority class **YES Cholera (1)** was poor. Oversampling was performed by using Adaptive Synthetic Sampling Approach (ADASYN), which is an improved version of Synthetic Minority Oversampling Technique (SMOTE) in order to restore sampling balance. ADASYN was selected because it can easily reduce the learning bias introduced by the original imbalance data distribution. It can also, adaptively shift the decision boundary towards the difficult to learn samples. In addition, ADASYN is independent of the underlying classifier and can be easily implemented (Dorn & Jobst, 2002). Furthermore, decomposition or dimensionality reduction of the dataset was performed using Principal Component Analysis (PCA). This is because; PCA can effectively reduce the high dimensionality of data by selecting an optimal feature from the original dataset (Lever, Krzywinski & Altman, 2017).

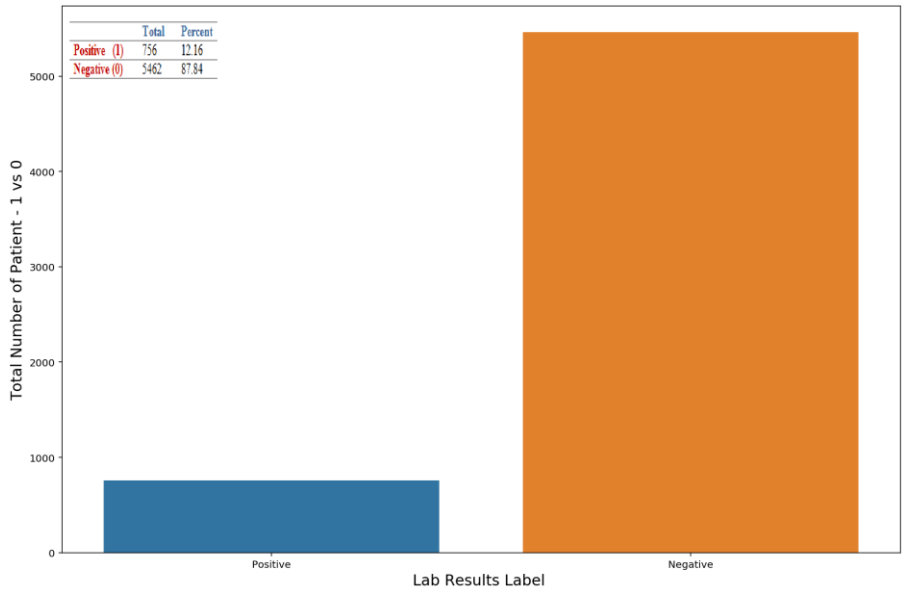


Figure 27: Imbalanced Cholera Dataset

3.9 Model Evaluation Metrics

Based on the nature of the collected cholera dataset, balanced-accuracy, sensitivity, specificity and F1-score metrics were used to evaluate the performance of the models, as shown in Table 8. The balanced-accuracy was performed on the dataset in order to avoid overvaluing the **NO Cholera (0)** label due to the number of present samples. A description of these metrics is provided in the next paragraph.

Specificity and sensitivity are clinometric parameters that together define effectively the presence or absence of specific conditions such as outbreak or diseases. The F1-score is a suitable measure of models tested with imbalance datasets; which is calculated as the weighted harmonic mean of precision and recall (Kay, Patel & Kientz, 2015). Precision is the proportion of positive results that truly are positive, while Recall is the ability of a model to identify positive results correctly. F1-score is in range of 0 and 1, that is, the larger the value of F1-score the better the performance. Sensitivity is the ability of a test to correctly classify an individual as diseased, while specificity is the ability of a test to correctly classify an individual as disease-free (Akosa, 2017). Table 8 presents the description of sensitivity and specificity metrics.

F1-Score= 2* ((Prec*Recall)/(Prec+Recall)) …….... (Equation 1)

Where:

- Prec = TP/ (TP+FP) a measure of a classifier exactness.
- Recall = TP/ (TP+FN) a measure of a classifier completeness.

- TP is True Positive, TN: True Negative, FN is False Negative and FP is False Positive.

Table 8: Description of Sensitivity and Specificity

	Disease present	**Disease absent**	**Total**
Test positive	a (TP)	b (FP)	all cases
Test negative	c (FN)	d (TN)	all non-cases
	all diseased	all non-diseased	all participants in the study
	Sensitivity= a/(a+c)	Specificity= d/(b+d)	

(Lalkhen & McCluskey, 2008)

3.10 Model Selection

The algorithms were listed in decreasing order based on balanced-accuracy, sensitivity, specificity and F1-score metrics, and the top 3 selected. In order to confirm the observed performance differences were indeed statistically significant, a Friedman test, a non-parametric version of ANOVA was used. Friedman test is a non-parametric test for testing the differences between several related samples or values. Non-parametric means that the test does not assume the data comes from a particular distribution. Friedman test is an alternative for repeated measures analysis of variances (ANOVA), which is used to check statistical significance in the difference of means of three or more independent groups (Singh, 2013). Basically, it is used in place of the ANOVA test when the distribution of the data is not known. Friedman test was selected for selecting the best model or models because it is more widely applicable compared to ANOVA and can accurately analyze several related samples (Bewick, Cheek & Ball, 2004).

3.11 Model Fine Tuning

Voting classifier was chosen to combine predictions of the top 3 selected models. This is because; it is one of the most straightforward ensemble learning techniques which can combine multiples models (Singh, 2013). Then, grid search was implemented to find the best parameter values in order to obtain the best performance of the voting classifier (Schaer, Müller & Depeursinge, 2016). Lastly, the four-evaluation metrics – F1-score, balanced-

accuracy, sensitivity and specificity – were performed in order to obtain the evaluation values or metrics for the voting classifier.

3.12 Model Evaluation

In this study, model evaluation was categorized into four processes. The first one was the model evaluation process using a comparative analysis theory whereby the process used the cholera datasets without the weather variables. The second one was the model evaluation process using the dataset datasets from other regions (Tanga and Songwe), the third one was the assessment of the model's impact and official acceptance to the study area and its stakeholders and the fourth model evaluation evaluated the performance the developed ML model in terms of its usability, extensibility (incorporating environmental variables) and computational performance compared to the existing cholera mathematical models. The following is their brief description;

The first model evaluation process was performed using comparative analysis theory. The aim of the first model evaluation validation process was to validate if weather variables have assisted in improving the prediction performance of the developed model. Its evaluation process used the same cholera dataset which was used in the model development. However, during this process the weather variables (rainfall, temperature, wind and humidity) were removed (cholera dataset without weather variables) in the cholera dataset in order validate whether the adding weather variables in the cholera prediction have assisted in improving the performance of the developed model. The evaluation process used the same model development approach as described in Section 3.6. In the process the dataset was run into the chosen 10 ML algorithms, then, the evaluation metrics were performed in order to select the best performing models or algorithms. Then, three best performing models were selected and then a model tuning process was performed in order to obtain the final one model known as voting classifier without weather variables. The first evaluation process was repeated ten times, on the same computer with the same internet speed and its average results were used as shown in Table 19. Lastly, the observation done by the 20 participants (which are described in Table 10) and Wilcoxon signed-rank test were used to confirm whether the integration of the weather variables added value to the model's performance. Wilcoxon signed-rank test is also a non-parametric statistical hypothesis which is used to compare two related samples (Singh, 2013).

The second model evaluation process was performed using a new set of datasets from Tanga and Songwe regions. The Tanga dataset had a total of 1179 patients while the Songwe dataset had 236 patients as shown in Fig. 29 and Fig. 30 respectively. The data were collected from TMA and the Ministry of Health and Social Welfare in Tanzania, which included a similar set of variables with the datasets used to develop and train the developed model, as shown in Table 5. The datasets for model validation were also pre-processed as explained in section 3.3. The main aim of this second model validation process is to check the robustness of the developed model through the use of other datasets with similar challenges (imbalance class etc.) and also, to validate if the codes can be referred and be adaptable enough to develop other models with new datasets. This is due to the fact that the developed ML predictive model has distinct features which are "adaptive" and "reference". Therefore, the datasets were just loaded into the codes in order to measure their results in terms of their performance in sensitivity, specificity, balance-accuracy and F1-Score.

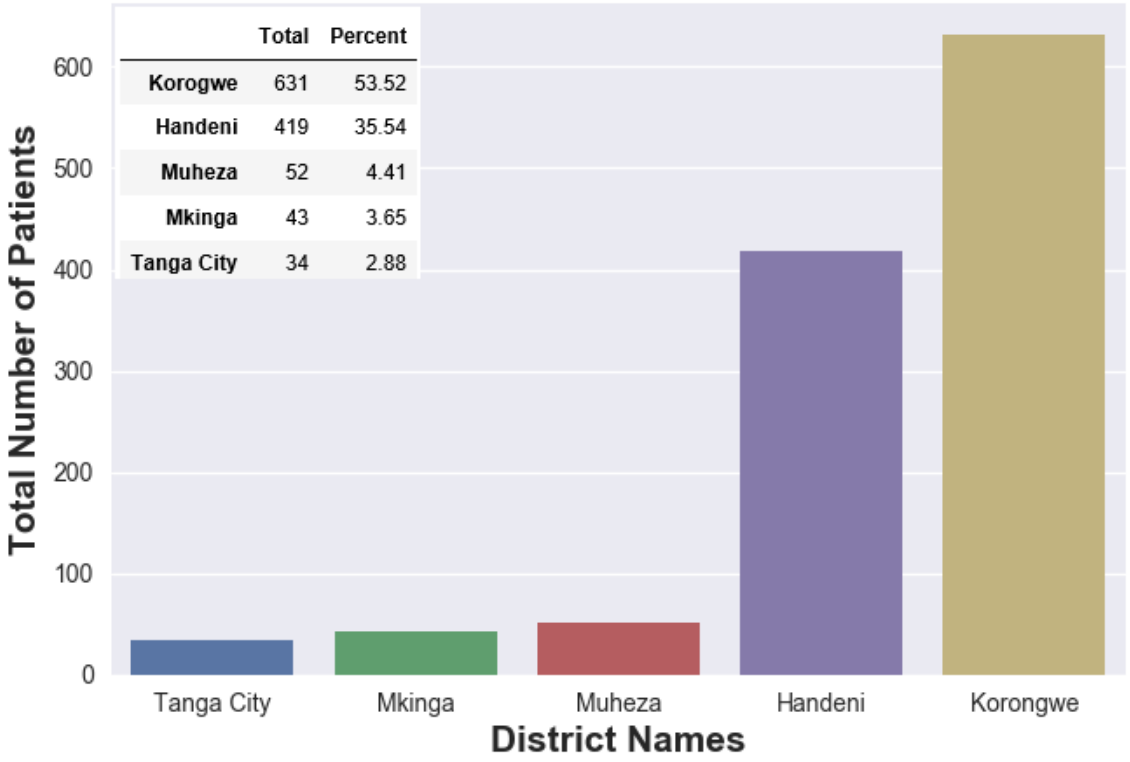


Figure 28: Patient Distribution in Tanga Region

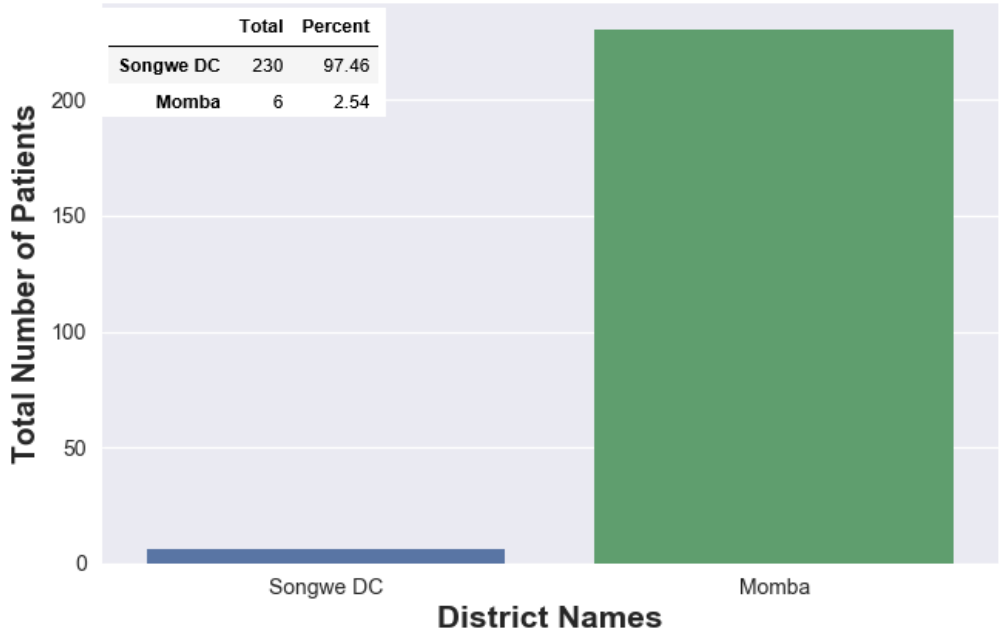


Figure 29: Patient Distribution in Songwe Region

The third model evaluation process was to assess the perspectives of health-care and environmental workers as well as cholera and diarrhoea patients on the feasibility, user-acceptance, performance, computational complexity, impact and awareness of using an integrated environmental factor-based healthcare predictive model to enhance effective cholera epidemics analysis and prediction in Tanzania. Ten research assistants were recruited to assist with the process. They were recruited through interviews, which were based on the knowledge of cholera management and health systems. Then, they were further trained on how to conduct the assessment using the developed interviewer-administered questionnaires. During the process, a mixed-design approach of quantitative and qualitative methods was employed as the study design. Focus group discussion and token were deployed in the process in order to get a large number of feedbacks and participants (500). Tokens ranging from 10 000 to 50 000 Tanzanian shillings was given to all participants. Whereas, interviewer-administered questionnaires were used especially to the medical officers in order to get deep understanding on the knowledge of cholera management and Focus group discussions and interviewer-administered questionnaires focused on the following topics; knowledge on cholera management and its interventions, ML models for cholera prediction, access to safe water and sanitation level to mention a few. The focus group discussions were limited to a maximum of fifteen participants in order to ensure equal participation. Audio recorders, mobile phones and laptops were used in recording, transcription and coding recording the discussions and interviews. Then, focus group discussions and interviewer-

administered questionnaires as shown in appendix III were employed and administered to five hundred (500) as described in Table 9. The participants include medical officers, epidemiological analysts, nurses, environmental experts, ICT experts and cholera patients from Dar es Salaam region in Tanzania. Dar es Salaam region was selected for this process because, mostly is affected by cholera outbreaks in Tanzania and also, the collected datasets reflected the region August 2015 to April 2016 cholera outbreak with a total of 14 608 cases and 228 death incidences in the Dar es Salaam region (WHO, 2015).

Table 9: Participants' Profile and Characteristics

Participants' Characteristics	Total Sample (N=500)	Percentage Distribution (%)	Description
Age in years			The average age group for the study was 28 years old.
▪ ≤ 20	85	17	
▪ 21-35	330	66	
▪ ≥ 36	85	17	
Occupation			Participants with features such as Trader, Artisan, Farmer, Salary Worker and Unemployed fall under Patient Cluster.
▪ Trader	50	10	
▪ Artisan	20	4	
▪ Farmer	60	12	
▪ Salary worker	30	6	
▪ Medical expert	200	40	
▪ Environmental expert	50	10	
▪ ICT expert	50	10	
▪ Unemployment	40	8	
Education level			Secondary consisted of participants with Junior and Senior high school levels at a ratio of 3:2, respectively. Non-Formal Education consisted with participants who do not have any kind of education.
▪ Primary	100	20	
▪ Secondary	110	22	
▪ Certificate Level	30	6	
▪ Diploma Level	40	8	
▪ University	200	40	
▪ Non-Formal Education	20	4	
Cluster level			Strictness was applied in selecting all participants based on five clusters. Therefore, there were no participants under the Non-Clustered category.
▪ Patient	200	40	
▪ Medical Experts	200	40	
▪ Environmental Experts	50	10	
▪ ICT Experts	50	10	
▪ Non-Clustered	0	0	
District			The study unbiasely considered all districts in Dar es Salaam regions. In addition, equal distribution of participant ration was highly considered.
▪ Kinondoni	100	20	
▪ Temeke	100	20	
▪ Ilala	100	20	
▪ Ubungo	100	20	
▪ Kigamboni	100	20	
Gender			Majority of participants were female at Female to Male ratio of 3:2.
▪ Female	60	60	
▪ Male	40	40	

Furthermore, data pre-processing and analysis was performed whereby the collected quantitative data was checked for presence of errors such as missing data, typos, misspellings, completeness and accuracy. Then, the qualitative data from voice recorders and phone audio recorders were also recorded thoroughly with a team of ten research assistants to ensure completeness and accuracy of the data. The two sets of data were integrated into one file, then repeatedly cleaned and analysed using Statistical Package for Social Science software (SPSS) (Miller *et al*., 2007). Descriptive statistics including frequency and cross tabulation on all variables were run to generate results of the study.

The fourth evaluation was to check on how well the ML model performs in terms of its usability, extensibility (incorporating environmental variables) and computational performance compared to the existing cholera mathematical models.

3.12.1 Usability

During this process, the focus group discussions and interviewer-administered questionnaires as shown in appendix IV were employed and administered to twenty (20) as described in Table 10. During the interview and group discussion questions related to comparison between Codeco Model and the developed ML Model were featured which include (a) which model is easier to understand? (b) which model is easier to use? (c) which model is user friendly? Etcetera.

3.12.2 Extensibility (Incorporating Environmental Variables)

During this process also, the 20 participants (as described in Table 10) were shown the process of incorporating the additional variables into the Codeco Model and the developed ML Model. Then, interview and group discussion questions related to the comparison between Codeco Model and the developed ML Model were featured which include (a) Which model is easier to incorporate additional variables? (b) Which model is fast to understand the procedure of integrating additional variables?

3.12.3 Computational Performance

In the current study, computation performance is measured as the elapsed time necessary to make prediction. During this process MATLAB used to test the computational performance when variables are added into the mathematical model equations. The programming language

MATLAB is a commercial programming language which has lost some market share (Paradis, 2005). Then, python was used to check also the computational performance when variables are added in the Both experiments were done on a laptop with the following specifications: Processor type: Intel Core i3, Clock speed: 2.1 GHz, Memory (RAM): 4 GB RAM, 320 GB HDD and Operating system: Window 10. In addition, internet speed when conducting the experiments were kept constant (dedicated bandwidth), to ensure similar or constant working conditions. Lastly, the 20 participants were also involved to observe the results and then provided their option regarding mathematical and ML models as to which have more computational performance than the other:

(i) For mathematical model; the process used the Leo Model. The Leo Model is the extension or expansion of Codeco Model, which has considered the addition of two parameters into the Codeco Model formulation. The Codeco Model is explained in Section 2.2 and described in Fig. 7. These two parameters are the rate at which human beings are recruited (n1), and the natural death rate for human beings (n2). All other parameters in Leo Model are the same as those in the Codeco Model, and thus their definitions can be found in Table 2. We, therefore, increased the number of variables and parameters of the original Codeco Model to get a new model by incorporating the two parameters as given in the given equations **X, Y** and **Z**. In this experiment, the Leo Model has been extended by adding other variables and parameters to obtain three different model equations.

The equations for the Leo Model are:

Without exposed class (E)……. X

$$\frac{dS}{dt} = n_1 H - n_2 S - a\lambda BS \quad \text{.....(A1)}$$

$$dI/dt = a\lambda BS - rI - n_2 I \quad \text{.....(A2)}$$

$$\frac{dB}{dt} = B(nb - mb) + eI \quad \text{.....(A3)}$$

With exposed class (E)……...Y

The sub-group S after infection progress and become latent to the disease at the rate a. After a few days, the sub group E becomes infected and capable of transmitting the disease at the rate a.

$$\frac{dS}{dt} = n_1 H - n_2 S - a\lambda BS \quad \text{..... (A4)}$$

$$\frac{dE}{dt} = a\lambda BS - (\alpha + \delta)E \quad \text{..... (A5)}$$

$$\frac{dI}{dt} = \alpha E - rI - n_2 I \quad \text{..... (A6)}$$

$$\frac{dB}{dt} = B(nb - mb) + eI \quad \text{..... (A7)}$$

With exposed and recovery……. Z

The sub-group S after infection progress and become latent to the disease at the rate after a few days, the sub group E become infected and capable of transmitting the disease at the rate α. Individuals in subgroup I if treated they recover and attain temporally immune and move to sub-group R at a rate π, otherwise they die either naturally at the rate n_2 or due to the disease at the rate r. The individuals in sub-group R attain temporary immunity then return and become susceptible again at the rate a otherwise they die naturally at the rate n_2. Note $\delta = n_2$.

$$\frac{dS}{dt} = n_1 H + \mu R - n_2 S - a\lambda BS \quad \text{..... (A8)}$$

$$\frac{dE}{dt} = a\lambda BS - (\alpha + \delta)E \quad \text{..... (A9)}$$

$$\frac{dI}{dt} = \alpha E - \pi I - rI - n_2 I \quad \text{..... (A10)}$$

$$\frac{dR}{dt} = \pi I - \mu R - n_2 R \quad \text{..... (A11)}$$

$$\frac{dB}{dt} = B(nb - mb) + eI \quad \text{..... (A12)}$$

(ii) For ML models; similar comparative analysis with the same ML models was conducted ten times as explained in which at this process performance in terms of time (prediction latency) was collected and also, the average time was computed.

3.12.4 Accuracy

Based on the fact that both mathematical and ML models have various strategies such as ensemble methods of increasing the accuracy of the model. Therefore, during this process the 20 participants briefly gave the benefits of each in terms of their accuracy as explained in Section 2.2 and Section 4.3 and then, asked to select the one with higher accuracy than the other.

Table 10: Participants' Profile and Characteristics for the Fourth Evaluation

Participants' Characteristics	Total Sample (N=20)	Percentage Distribution (%)	Description
Age in years			-The average age group for the study was 30 years old.
▪ **≤ 20**	5	25	
▪ **21-35**	10	50	
▪ **□ 36**	5	25	
Occupation			-The process had equal participants distribution in terms of their profession-Cluster.
▪ **Medical expert**	5	25	
▪ **ICT expert**	5	25	
▪ **Mathematicians**	5	25	
▪ **Epidemiologists**	5	25	
Education level			-3 ICT experts and 1 Epidemiologist had Diploma level. Whereas University level consisted of Bachelor, Masters and PhD levels at the ratio 5:10:5 respectively.
▪ **Diploma Level**	5	25	
▪ **University**	15	75	
Location			-The study unbiasedly considered experts from different countries
▪ **Arusha, Tanzania**	5	25	
▪ **Dar es Salaam**	5	25	
▪ **Dodoma, Tanzania**	3	15	
▪ **Canada**	2	10	
▪ **USA**	2	10	
▪ **German**	3	15	
Gender			-We had an equal male to female distribution ration of 1:1.
▪ **Female**	10	50	
▪ **Male**	10	50	

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results of Climatology-aware HMIS

The section presents the results obtained towards the development of an effective Climatology-aware HMIS.

4.1.1 Proposed Functional and Non-functional Requirements

The result of the open-ended interview proposed the following functional and non-functional requirements for the proposed system as described in Table 11.

Table 11: Proposed System Requirements

Functional Requirements		
SN	**Requirements**	**Description**
1.	Submit health and environmental data, including the climatology variables, date and time on set, patient's personal, and home geographical location.	Users, Staging DB and EMIS should submit to and from, all essential data automatically.
2.	Search records per keywords and images.	The system should enable users to search for data using keywords and images.
3.	Generate visual reports.	The system should generate visual reports such as; excel and map-based reports which contain patient's details, laboratory results, geographical location, climatology variables, date and time on set.
		Non-functional Requirements
1	Security	The system should allow only registered users to access their information.
2	Efficiency	The system should do tasks required successfully without wasting time or energy.
3	Accessibility	The system should be easy to use, access, and allow offline registration and data submission when there is no internet connection.
4	Maintainability	The maintainability should easily be done by the system administrator.
5	Interoperability	The system should allow upgrading when needed.

4.1.2 Result of System Design Process

Based on the obtained system requirements, we adopted a client-server which utilizes a model view controller (MVC) domain. This is because; MVC can easily design, define the structure and enable rapid development of the proposed system. In addition, MVC facilitates code reuse and parallel development to enable modification of other parts of the system

without affecting the entire system (Leff & Rayfield, 2001) as shown in Fig. 31. In Fig. 31, the **CLIENT** module describes users of the system such as patients and system administrator. Then, the clients (users) use the **VIEW** module to send and receive the required information to and from the server. In return, the **CONTROLLER** module controls information flow from users to the staging DB, EMIS and the HMIS and vice versa.

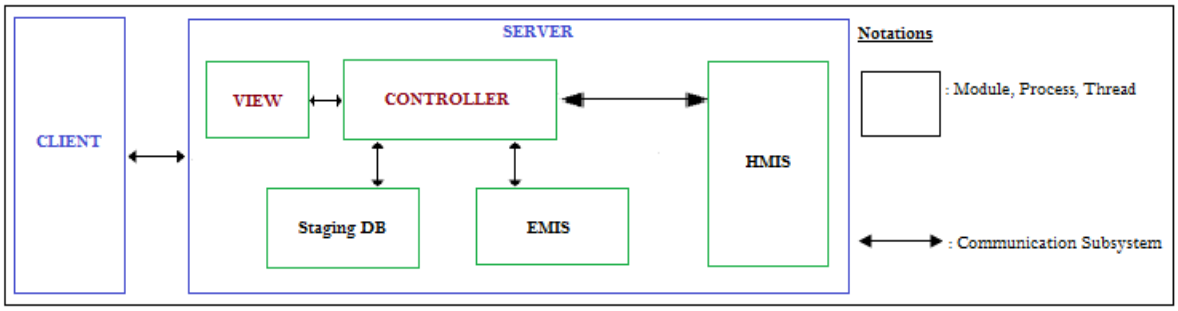


Figure 30: Design Model of the Proposed System

The model was further instantiated into a conceptual framework in order to provide a blueprint for development processes. In brief, the conceptual framework consists of a mobile phone application which enables the patients to securely register to the system, and then data is collected and sent to the staging database (Staging DB). The system administrator assists in the complete submission of environmental data including the climatology factors into the EMIS. Then, the communication subsystem integrates data from Staging DB and EMIS, and then sends them into the HMIS. At this stage, the data is ready to be sent to the registrar for further medical activities. Then the whole collected data and laboratory results of the patients are sent to the epidemiological analysts for further analyses.

4.1.3 Results of Feature Selection

This subsection describes the developed system of climatology-aware HIS, as shown in Fig. 32. At the development phase, the Staging, EMIS and HIS databases should be developed rapidly and in parallel in order to allow users to input and access information during the testing phase. Web interfaces which are GIS-based HMIS, and satellite, and sensor based EMIS should be developed using Java Server Page (JSP), JavaScript and Java codes to pull the collected data. Lastly, the mobile-app should also be developed using Java and connected to the MySQL database using JSON. The system should consist of the following:

(i) Mobile application, which is responsible for data collection and sensing, data processing such as; privacy preserving and feature extraction, and uploading data automatically to the Staging DB for storage.

(ii) EMIS, which is responsible for the collection of environmental data such as climatology variables, using sensors and satellite. The satellite and sensors work together to collect daily weather data, the load balancer cross-check the collected data before they are submitted to the EMIS.

(iii) Communication Subsystem which integrates the data from Staging Db and EMIS and then, submits them to the web-based HMIS for further processing.

(iv) Web-based HMIS which process the integrated data with lab-result, patient's historical data, and doctor's comments, and then, it transforms the data into required visual formats such as; excel for predictive model generation and map-based reports. The following are the main features of the developed system; data format interface, report sample of the system and security interface, which are described in subsection 3.3.1.

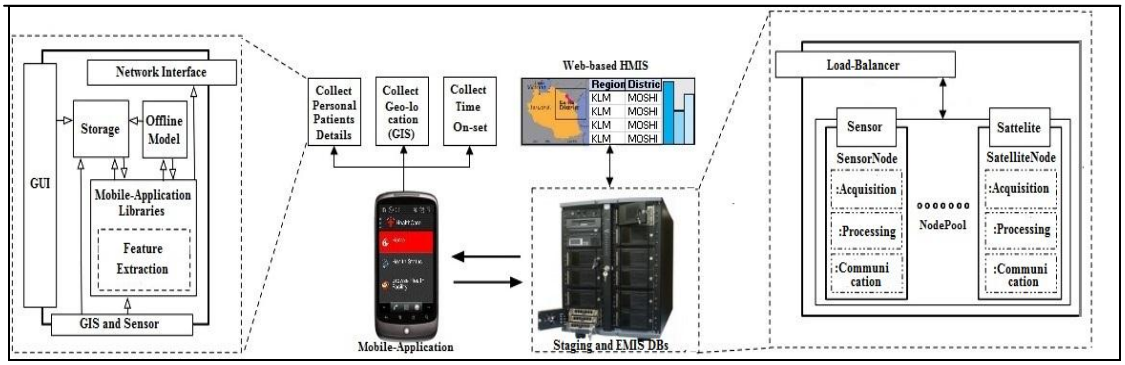


Figure 31: Description of the Developed System

4.1.4 The Data Format Interface

Figure 33 shows the required data format which includes patients' details with their environmental factors such as climatology aspects.

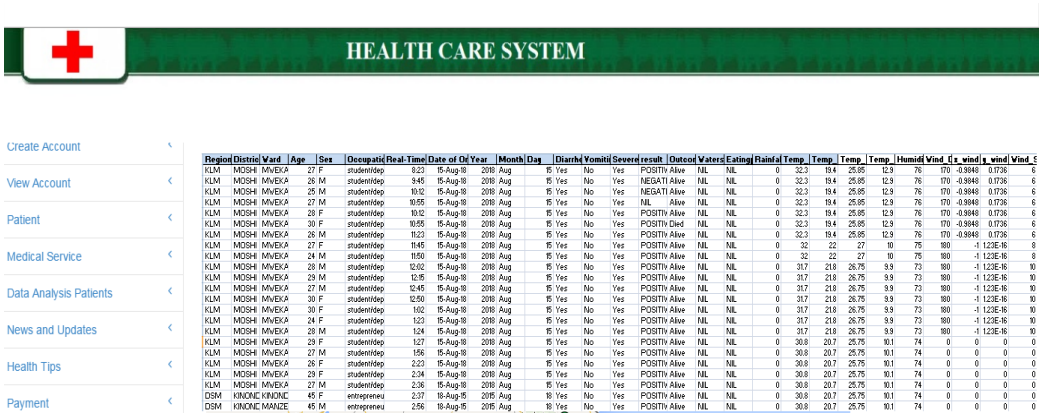


Figure 32: Data Format Interface of the Developed System

4.1.5 Result of System Evaluation

This subsection presents the evaluation results on whether the developed system truly does make it easier to manage Cholera epidemics.

(i) System Requirement Evaluation

Table 12 describes in brief, the fulfillment of the functional requirements in the developed system.

Table 12: System Requirements Verification Featured in the Questionnaires

Requirements	System Verification
The developed system should collect health and environmental data, including the climatology variables, date and time on set, patient's personal, and home geographical location on-time.	The developed system has integrated advanced ICT to collect the required data for the analysis into one platform hence, timely data collection and analysis is inevitable.
The developed system must be able to produce required data format and store data acquired.	The developed system consists of Staging DB, EMIS and communication subsystem to manage and control the required data format and storage.
The developed system must be able to manipulate the collected data and establish required data analysis report format.	The developed system has integrated advanced ICT such as mobile application and GIS in the control module in order to analyze data on-time.
The developed system should secure data effectively.	This is achieved through the adoption of client-server architecture style across all modules or parts of the developed system. The architecture style enables users to access the system as clients by sending requests to the server while the information control subsystem is associated with a back-end server to handle requests from clients. In addition, all patients have a unique password and identification for accessing the system (mobile or web application).

Requirements	System Verification
The developed system must be simple and flexible.	The developed system adopts MVC and RAD which iteratively break the task into small modules, then create them independently and then join them together to represent a larger system. The modular design facilitates customization by allowing different parts of the system to be easily added, removed, and rearranged. Hence the developed system is flexible, simple and can easily be customized to fit the needs of the users.

(ii) System Evaluation

The results of system evaluation showed that at an average of 87% provided positive comments on the developed system as described in Fig. 34. The remaining 13% of the results, the majority of participants were not sure of getting an affordable internet connection. Hence, we trained the participants on how to easily and cheaply connect their Smartphone with reliable internet connections in Tanzania. In addition, the analytical result of the system showed that 88.5% of participants (households) had Cholera disease.

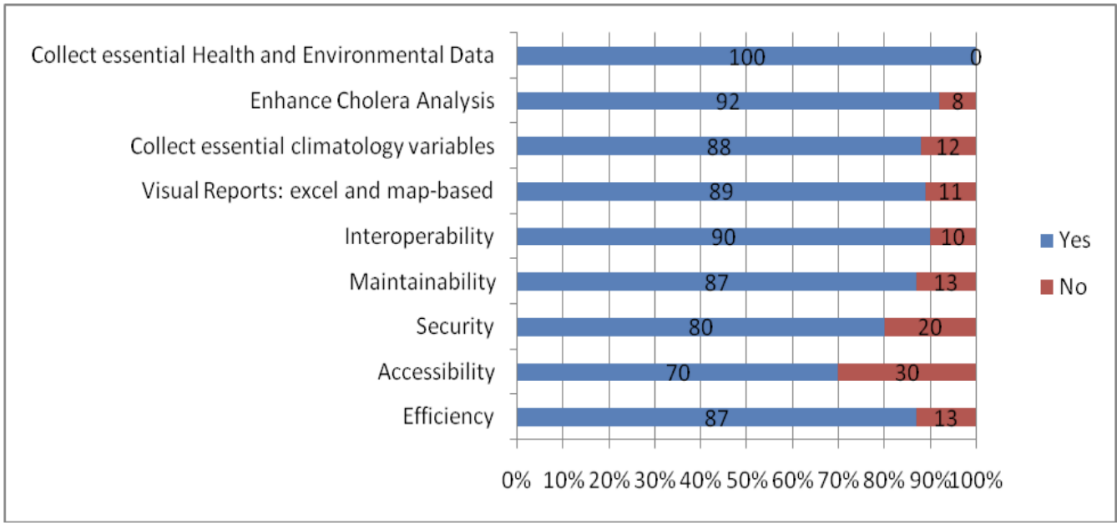


Figure 33: Results of System Evaluation

4.2 Results of Data Collection Requirements

As noted in the literature review section that, the *V. cholerae* organism thrives in the aquatic environment and most researchers have noted that high air temperatures and periods of excessive rainfall create environmental conditions that favor the bacteria's growth (Lugomela *et al*., 2015). Additionally, in dry conditions, river levels decrease and bacteria accumulate in dangerously high concentration. On the other hand, during excessive rainfall, flooding can

spread bacteria to regions that had not previously been infected, resulting in fast-spreading epidemics (Kuhn *et al*., 2005). Researchers further anticipate the rise in the number of cholera outbreaks in the future as a result of climate change (Kuhn *et al*., 2004). Thus, to enhance cholera prediction, the weather variables were included in the data requirements or characterization for the development of the proposed model. Given the collected datasets, a features selection process was performed. Moreover, the literature proven theories as shown in Table 13, together with the results of feature selection and outliers were used to select the best performing variables for cholera prediction.

Table 13: The Influence of Environmental Factors on Cholera and *Vibrio Cholerae*

Factors	Climate and Weather Drivers	Influences
Temperature	Seasons, inter-annual variability.	The growth of *Vibrio cholerae*, plankton's productivity and infection by temperate phases.
Rainfall	Seasons, precipitation, monsoons, sea level rise, atmospheric decomposition and salinity.	The growth of cholera and expression of cholera toxin.
Humidity	Seasons and inter-annual variability in light and nutrients.	The growth of *Vibrio cholerae*, attachment to planktons and survival of *Vibrio cholerae*.
Wind	Seasons, inter-annual variability.	The spread of cholera
Salinity	Seasons, monsoons and sea level rise.	The growth of *Vibrio cholerae* and seroconversion expression of cholera toxin.
Ph	Seasons and inter-annual variability.	The growth of *Vibrio cholerae*
Fe3+	Precipitation, atmospheric deposition	The growth of *Vibrio cholerae* and expression of cholera toxin.
Exogenous products of algal growth and chitin	Seasons, monsoons and inter-annual variability in light, nutrients and planktons.	The growth of *Vibrio cholerae*, survival of *Vibrio cholerae* and attachment to exoskeletons.
Human being or human behaviour	In most cases, cholera is transmitted from one person to another and its occurrence is mostly seasonal. In the same way, human behaviour or activities change seasonally in relation to available weather (Shuai & Van den Driessche, 2015). For example, in drought seasons, most people are less careful on cleanliness in using urinary facilities, cooking and overall hygiene. It is during this time that human behaviour leads to the creation of favourable conditions for the growth of *V. cholerae* in the surrounding areas. If such areas experience heavy rains, cholera infection is likely to occur, as people sometimes fetch rainwater for cooking. Moreover, heavy rainfall can spread *V. cholerae* from one point to another (Rebaudet *et al.,* 2013). Movement of the infected population can also lead to long-range dissemination of *V. cholerae* to far-	Cholera epidemics are also affected by seasonality in human behaviour in terms of their mobility and daily activities (Mari *et al*., 2012). Moreover, such factors as human living location and sanitation, level of knowledge on cholera disease, age factor and economy can also lead to the occurrence and spread of cholera. For instance, human activities such as poor methods of fishing and farming may impact ecological balances, which may potentially lead to such diseases as cholera, which is associated with environmental changes like floods and droughts (Soto, 2009).

Factors	Climate and Weather Drivers	Influences
	reaching water bodies. Most of these major movements are done seasonally. For example, most people who live in areas prone to flooding move to other areas for their wellbeing. Other movements occur during such seasons as school opening and closing, farming and fishing seasons, to mention a few (Medsker *et al.*, 2016). Therefore, given that human mobility behaviour greatly influences the emergence of diseases far from epidemics hotbeds, it is then being a major determinant of cholera epidemic spread and prediction. In fact, this has been a key issue in the dynamics of any infectious disease. However, it has received relatively low attention with specific reference to cholera epidemics (Medsker *et al*., 2016).	

(Lipp, Huq & Colwell, 2002)

4.2.1 Results of Feature Selection

Given the above explanation of how different factors can influence the prediction of cholera epidemics. This section presents how an intensive feature selection was performed in identifying key features or a set of datasets used in model development. In this process RF algorithm was used to perform feature selection because of its robustness. Fig. 35 below shows the approach used in feature selection, while Table 14 shows the selected features for the model development (result of feature selection).

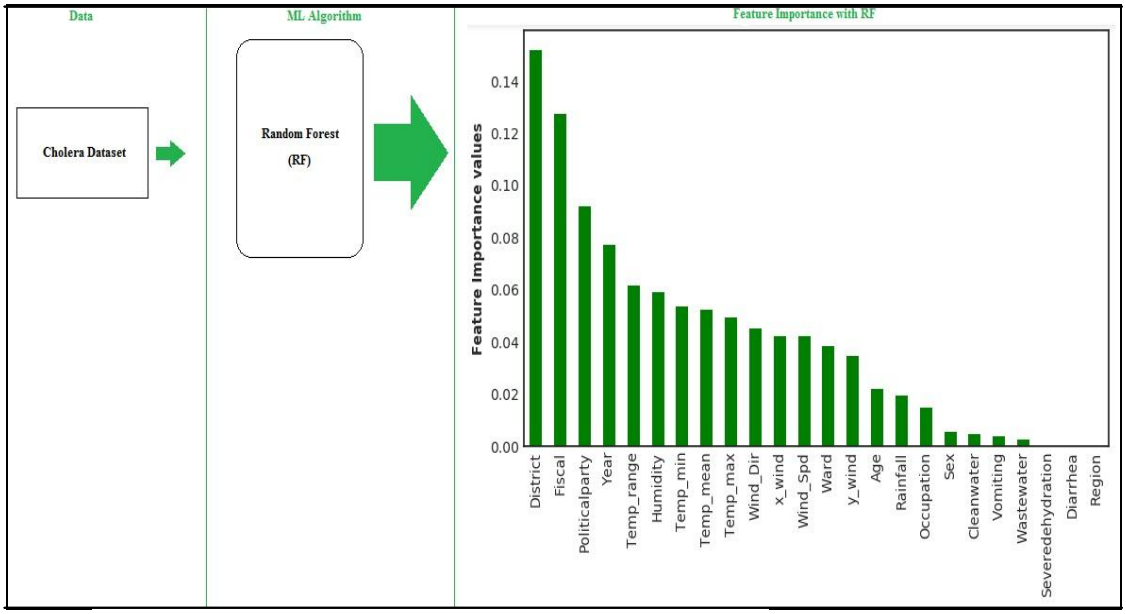


Figure 34: Feature Selection Approach Using RF Classifier

Table 14: Selected Features for Model Development

Variables	Description	Measurements of Variables
Daily Seasonal Weather Changes Data		
Temp_mean	Mean Temperature	Degree Celsius (°C)
Rainfall	Rainfall	Millimetre (mm)
Humidity	Relative Humidity	(%)
Wind_Dir	Wind Direction	Degrees
Cholera Cases Data with Regards to the Patient Details		
District	District location	Dar es Salaam Districts
DOS	Date on set	Date-Month-Year
LabResults	Laboratory Result	Positive or Negative
Water Source Type		
WasteWater	TreatedWater/CleanWater (1) and UntreatedWater / DirtyWater (0)	1 and 0

4.3 Results of Model Development

4.3.1 Imbalance Data Challenge

In order to handle the imbalance class, the ADASYN sampling techniques assisted to restore sampling balance by reducing the learning bias to 0.502. Thereafter, PCA was applied to reduce the high dimensionality of data by selecting an optimal feature from the original dataset. In addition, the study adopted n-fold cross validation in order to generalize predictive ability and avoid over-fitting problem which may arise while developing a model. Then, the

model building process commenced using the 10 best performing algorithms for cholera prediction.

4.3.2 Model Evaluation Metrics

Table 15 shows the results of the used three evaluation metrics, namely F1-score, sensitivity, and balanced-accuracy metrics. For good visualization purpose the following Fig. 36, 37, 38 and 39 present histograms for the results of F1-score, sensitivity and balanced-accuracy metrics.

Table 15: Results of F1-score, Sensitivity, Specificity and Balanced-Accuracy Metrics for Ten Algorithms

SN	Model Name	Performance (F1 Score)	Sensitivity	Balanced Accuracy
1.	XGBoost Classifier	0.210	0.994	0.558
2.	Gradient Boosting Classifier	0.304	0.995	0.590
3.	Random Forest Classifier	0.610	0.982	0.740
4.	Bagging Classifier	0.613	0.994	0.741
5.	MLP Classifier	0.346	0.988	0.608
6.	K-NN	0.320	0.989	0.597
7.	Decision Tree Classifier	0.456	0.985	0.656
8.	Supper Vector Classifier	0.090	0.999	0.523
9.	ExtraTree Classifier	0.590	0.996	0.719
10.	Logistic Regression Classifier	0.057	0.989	0.513

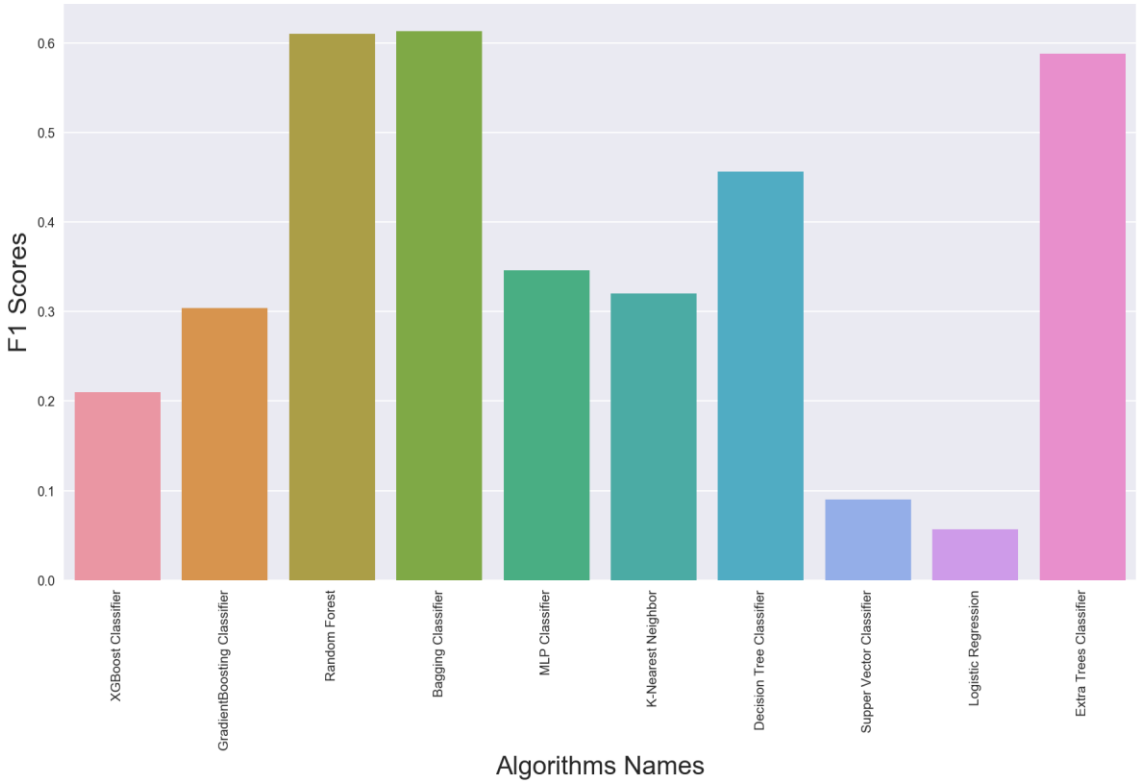


Figure 35: Results of F1_Score Metrics for Ten Algorithms

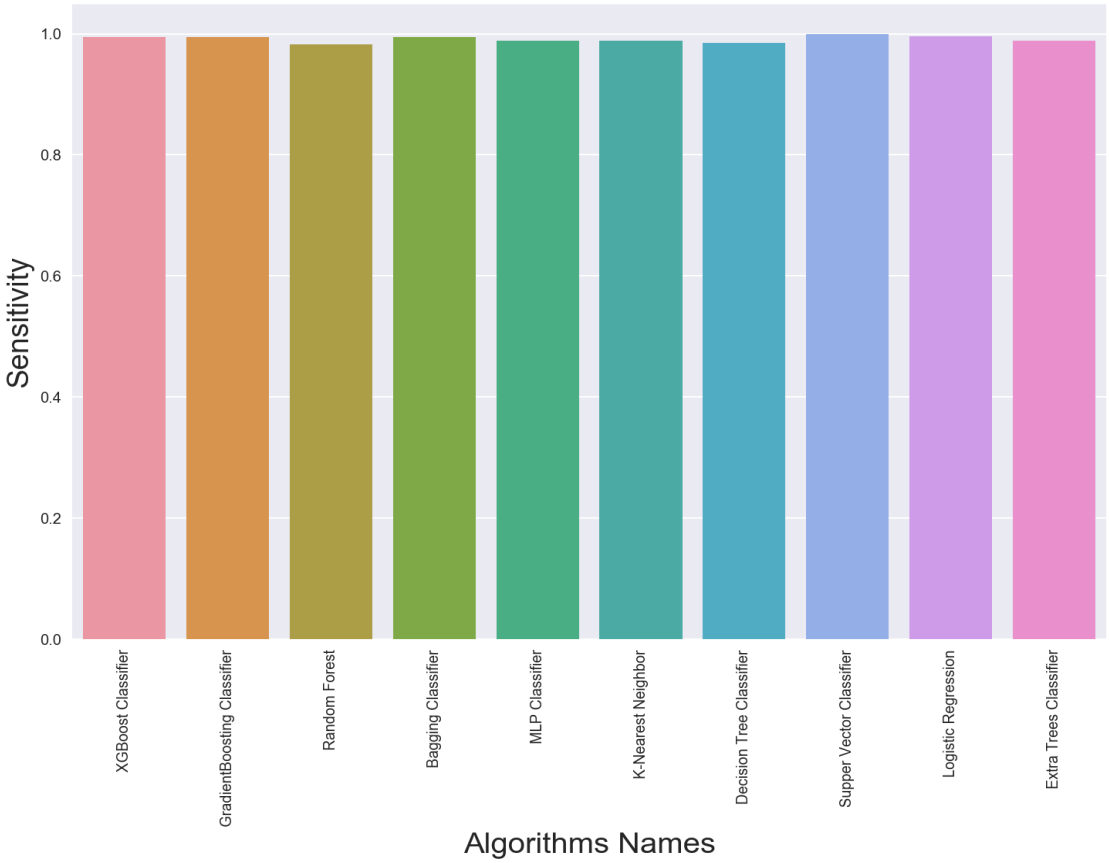


Figure 36: Results of Sensitivity Metrics for Ten Algorithms

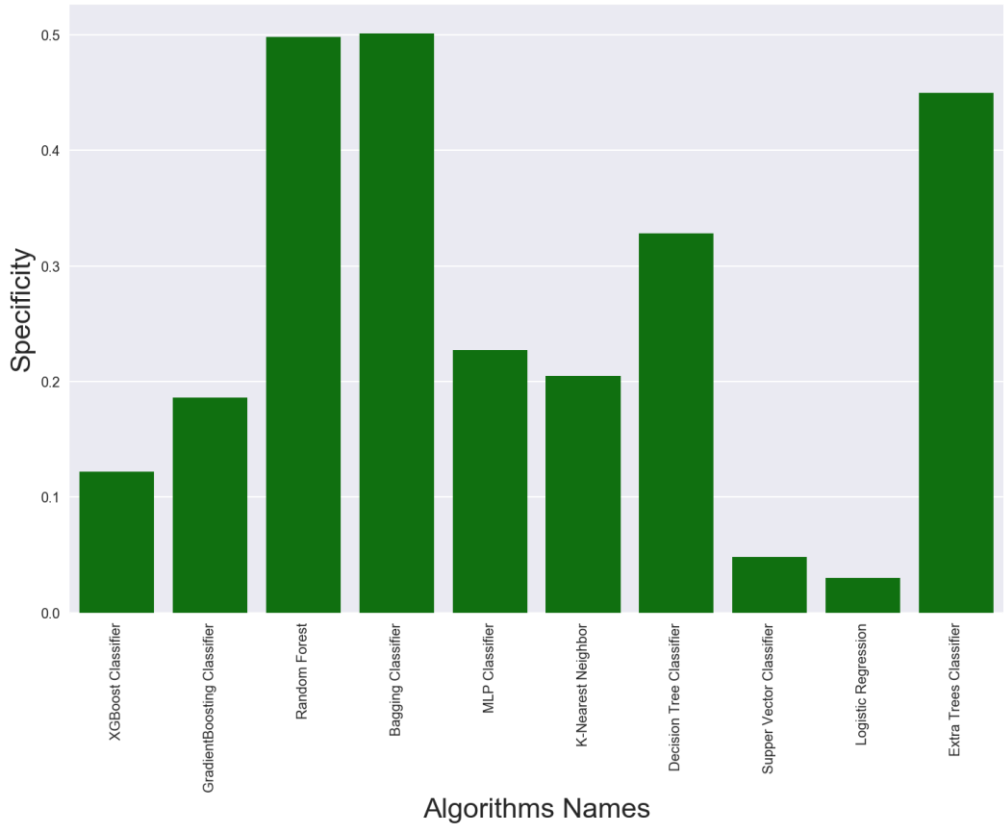


Figure 37: Results of Specificity Metrics for Ten Algorithms

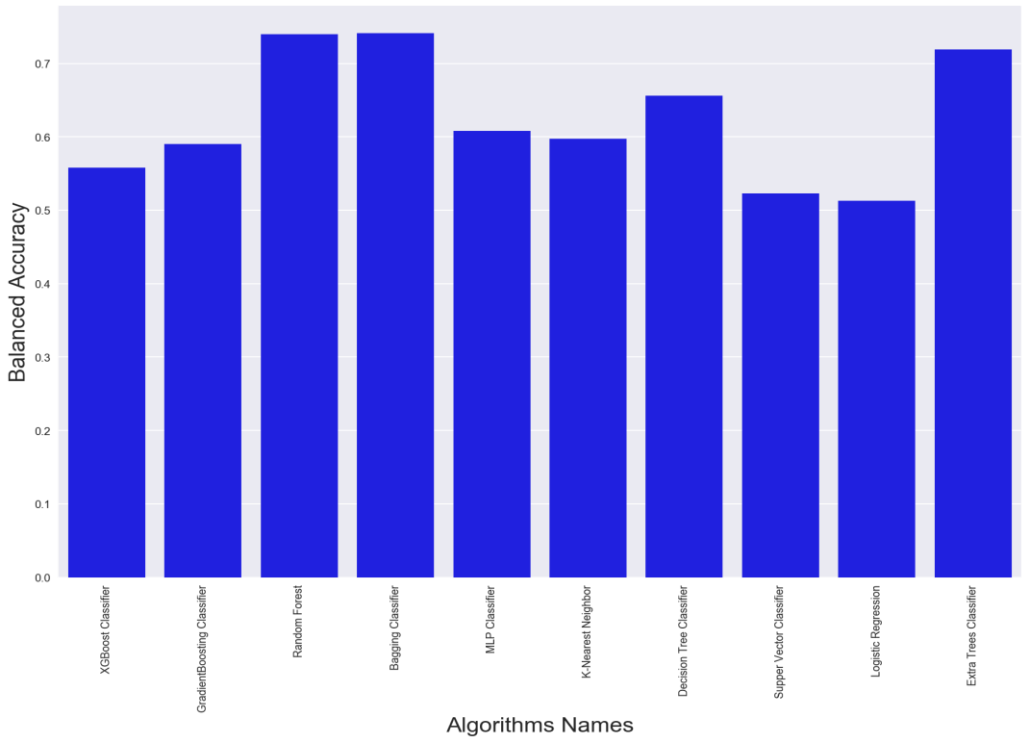


Figure 38: Results of Balanced-Accuracy Metrics for Ten Algorithms

4.3.3 Model Selection

Based on physical observation on the evaluation metrics results, the nature of the research study, characteristics of each algorithm and the results of Friedman (FM) test rank as shown in Table 4.6, three models with the best results were obtained. Moreover, the result values of F1-Score were mostly observed. This is because it is a good evaluation metric to evaluate the performance of a model when working on an imbalance dataset. It also provides a clear picture on how the model predicts each class. Table 16 shows the results of evaluation metrics for the three best performing algorithms, which are F1-score, sensitivity and balanced-accuracy metrics. For, good visualization purposes, Fig. 40, 41 and 42 present histograms for the results of F1-score, sensitivity, specificity and balanced-accuracy metrics respectively.

Table 16: Results of FM Test Rank for the Best Ten Algorithms

Model Name	XGBoost	GB	RF	Bagging	MLP	K-NN	DT	SVM	ET	LR
FM Results	5	4	2	1	6	7	8	9	3	10

Table 17: Results of F1-Score, Sensitivity, Specificity and Balanced-Accuracy Metrics for the Best Three Algorithms

SN	**Model Name**	**Performance (F1 Score)**	**Sensitivity**	**Balanced Accuracy**
1.	RF Classifier	0.610	0.982	0.740
2.	Bagging Classifier	0.613	0.994	0.741
3.	ExtraTree Classifier	0.590	0.996	0.719

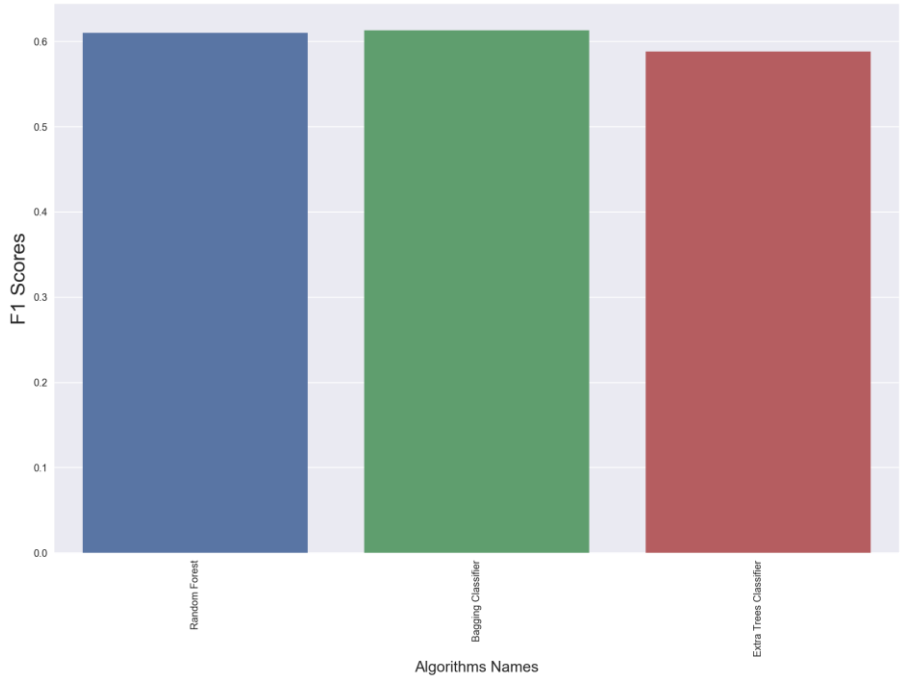


Figure 39: Results of F1_Score Metrics for the Best Three Algorithms

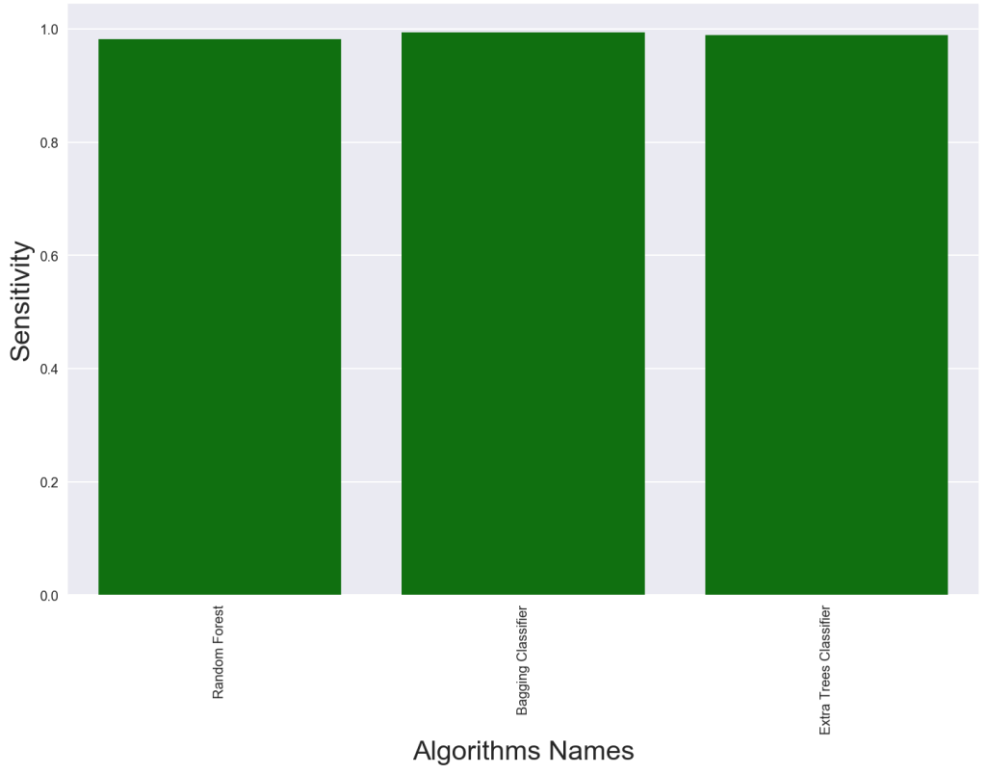


Figure 40: Results of Sensitivity Metrics for the Best Three Algorithms

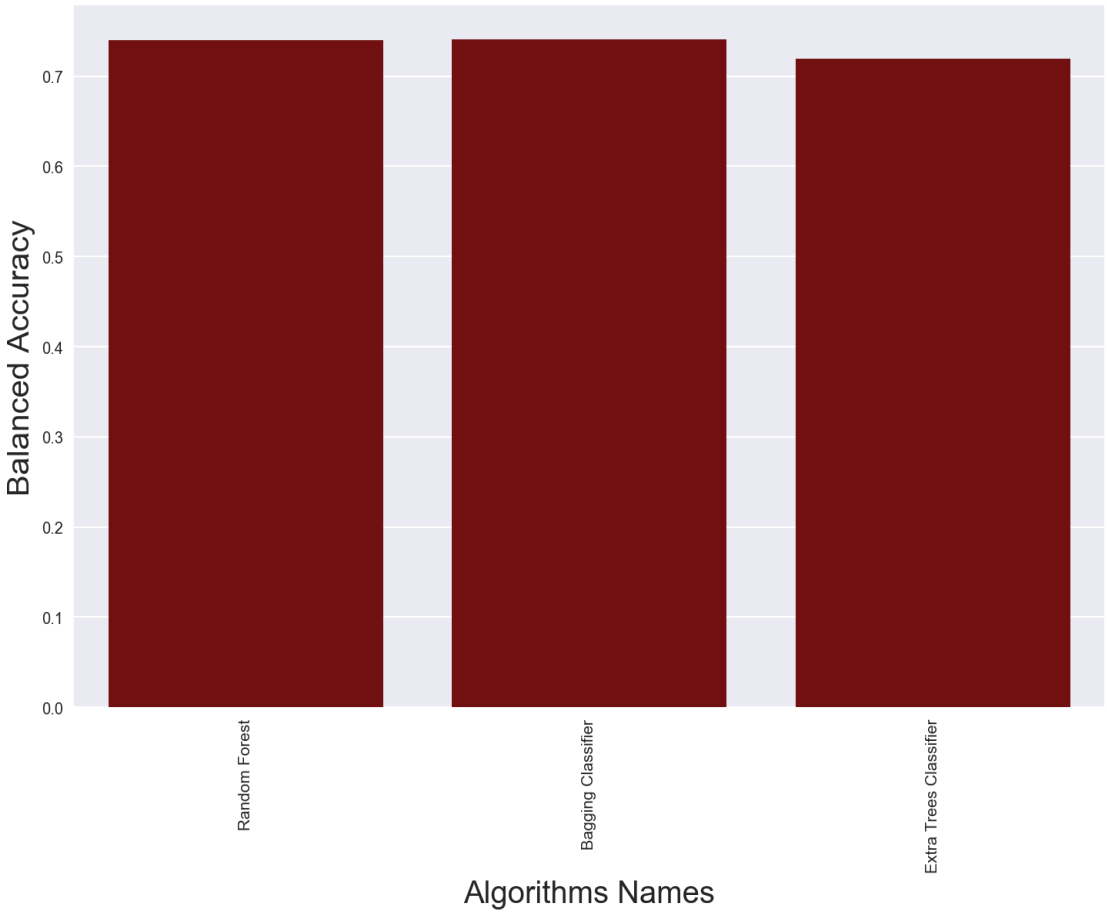


Figure 41: Results of Balanced-Accuracy Metrics for the Best Three Algorithms

4.3.4 Model Fine Tuning

In addition, we performed a model fine tuning process in order to obtain one best performing model from the three select the best three models. In model, fine tuning ensemble method was deployed in order to improve the results by combining the performance of the best three performing models; which lead us into obtaining our final model result called voting classifier whose evaluation metric results are shown in Table 18 and Fig. 43.

Table 18: Results of F1-Score, Sensitivity, Specificity and Balanced-Accuracy Metrics for Voting Classifier

SN	Model Name	Performance (F1 Score)	Sensitivity	Balanced Accuracy
1.	Voting Classifier	0.651	0.965	0.785

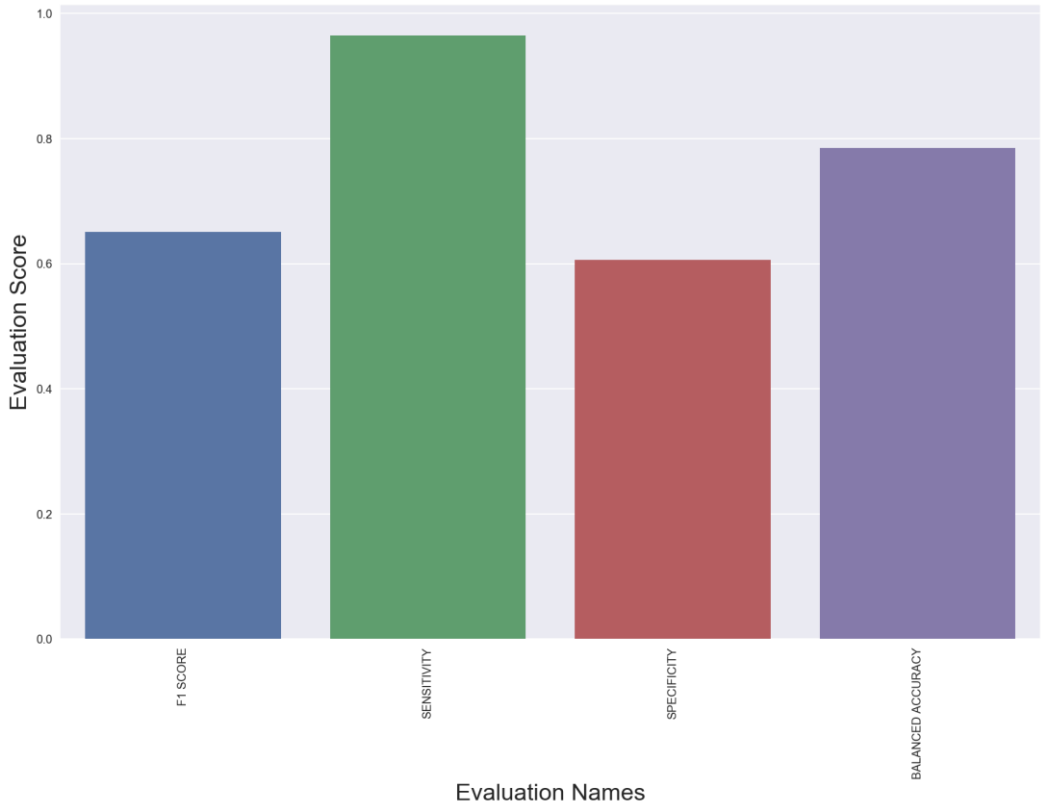


Figure 42: Evaluation Metric Results for Voting Classifier

4.3.5 Results of Analytical of Cholera Epidemics based-on Seasonal Weather Variables

Based on the analysis process on cholera dataset, we were able to observe that temperature ranges from 22^{o}C to 32^{o}C, the rainfall level higher than 50 millimetres and the humidity level higher than 75% are the favourable conditions for cholera incidences. The model failed to predict precisely in terms of time and specific weather variables due to the following data limitations and challenges which include:

(i) The cholera dataset had a lot of data-inconsistency.

(ii) Lack of enough cholera dataset for the process.

4.4 Results of Evaluation of Model's Performance

The results of evaluation of the developed model's performance are categorized into four processes. The first one is on the model evaluation process using the same set of cholera datasets but without the weather variables. The second model evaluation used new cholera datasets from Tanga and Songwe regions. The third model evaluation process, involves the

assessment of the overall research study's impact and its official acceptance to the study area and its stakeholders. Lastly, the fourth model evaluation process.

4.4.1 Results of Model Performance using the Cholera Dataset without Weather Variables

After running experiments without the weather variables, it was observed that there is a drop in the values of evaluation metrics especially in F1-Score, Sensitivity and Balance-Accuracy of the ten algorithms. In addition, the process led into obtaining new best three performing algorithms which are RF, XGB and KNN. Then, after computing for the voting classifiers with the best selected three algorithms (RF, XGB and KNN) and the best selected three algorithms which were obtained in the model with weather variables which are RF, BG and ExtraTree; both results showed drop down of the evaluation metric values as shown in Table 19 and Table 20 respectively. Based on this observation done by the 20 participants (which are described in Table 10) and Friedman test, the evaluation study as shown in Table 20, Table 21 and Fig. 44 confirms that it is important to involve or integrate the weather variables in the prediction of cholera epidemics. This is because they assist to boost the analysis and prediction accuracy of cholera models and systems in general.

Table 19: Results of F1-score, Sensitivity and Balanced-Accuracy Metrics for Ten Algorithms_Without Weather Variables

SN	Model Name	Performance (F1 Score)	Sensitivity	Balanced Accuracy
1.	XGBoost Classifier	0.411	0.966	0.642
2.	Gradient Boosting Classifier	0.372	0.988	0.617
3.	Random Forest Classifier	0.409	0.958	0.643
4.	Bagging Classifier	0.400	0.952	0.640
5.	MLP Classifier	0.048	0.998	0.511
6.	K-NN	0.411	0.963	0.642
7.	Decision Tree Classifier	0.371	0.911	0.634
8.	Supper Vector Classifier	0.220	1.000	0.562
9.	ExtraTree Classifier	0.375	0.952	0.627
10.	Logistic Regression Classifier	0.007	0.598	0.013

Table 20: Results of F1-Score, Sensitivity, Specificity and Balanced-Accuracy Metrics for Voting Classifier_Without Weather Variables

SN	Model Name	Performance (F1 Score)	Sensitivity	Specificity	Balanced Accuracy
1.	Voting Classifier	0.417	0.959	0.333	0.646

Table 21: Results of F1-Score, Sensitivity, Specificity and Balanced-Accuracy Metrics for Voting Classifier_Without Weather Variables

SN	Model Name	Performance (F1 Score)	Sensitivity	Specificity	Balanced Accuracy
1.	Voting Classifier	0.377	0.969	0.278	0.624

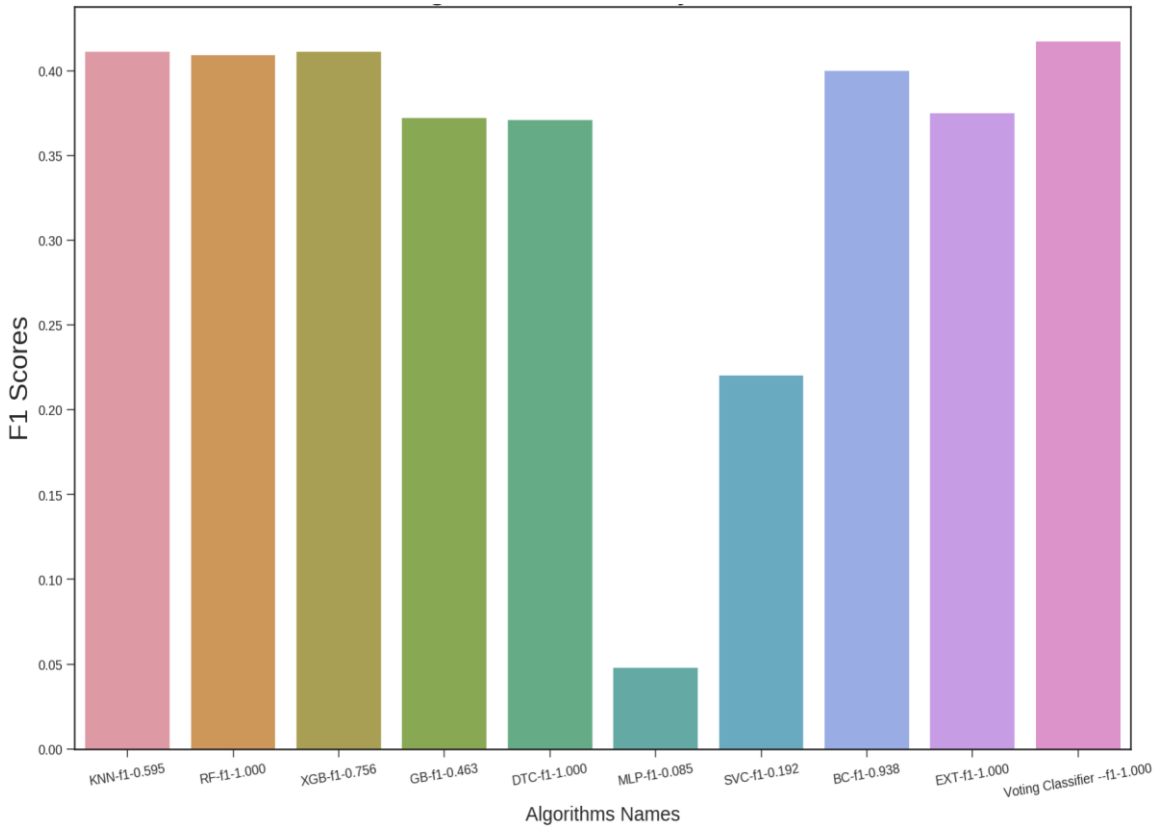


Figure 43: Results of F1_Score Metrics for Ten Algorithms and Voting Classifier_Without the Weather Variables

4.4.2 Results of Model Performance Using New Datasets from Different Regions

The evaluation process used the developed voting classifier in order to test and validate its robustness using datasets from Tanga and Songwe regions. Figure 45, Table 18, Fig. 46 and Table 19 show the patients' distribution according to LabResults (positive and negative) for Tanga region, results of F1-score, sensitivity, specificity and balanced-accuracy metrics for Tanga-voting classifier, patients' distribution according to LabResults (positive and negative) for Songwe region and results of F1-score, sensitivity, specificity and balanced-accuracy metrics for Songwe-voting classifier respectively. However, the models were able to predict well if the person has cholera but it predicts poorly if the person is not affected by cholera due to the fact that the number of Tanga and Songwe datasets were fewer compared to the Dar es Salaam dataset. In other words, you can say that the model can predict well in one class (YES Cholera (1)) and poor in another class (NO Cholera (0)). Based on the results shown in Table 22 and Table 23, it is clear that the selected model is robust enough at an average rate of 69.05% to handle the prediction of cholera with linkage to weather changes in Tanzania's settings. However, the results also led into recommending the need to collect quality datasets and also, restructure model parameters and selection of the best three performing classifier when dealing with a new set of datasets. This is because when the new set of data was used it was observed that there is a decrease of model's accuracy due to poor quality of data from Tanga and Songwe regions.

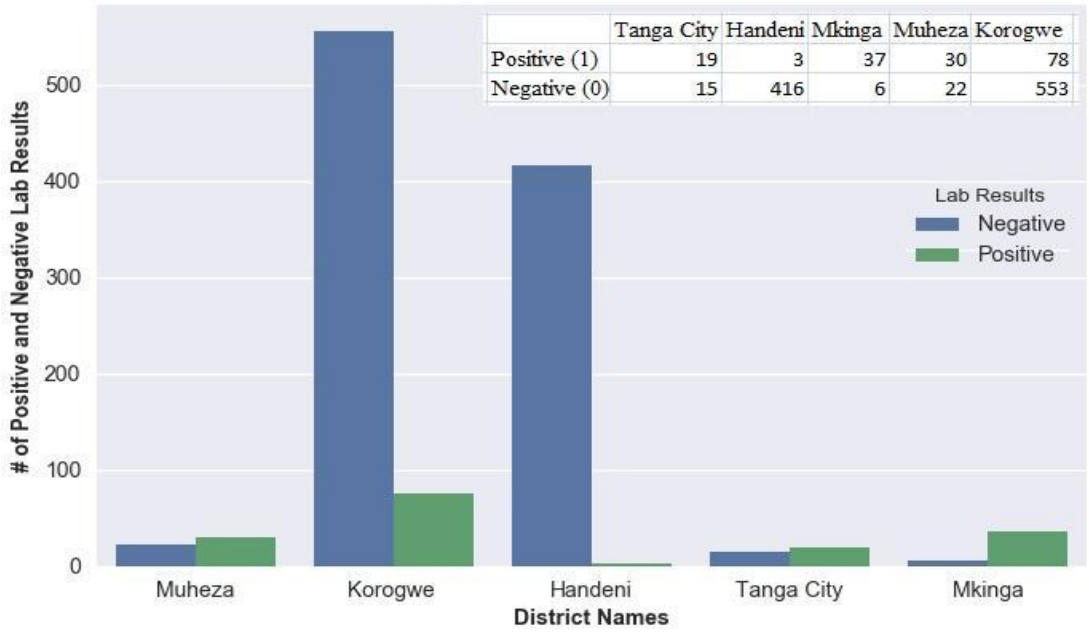


Figure 44: Patients Distribution as per LabResults (Positive and Negative) in Tanga

Table 22: Results of F1-Score, Sensitivity, Specificity and Balanced-Accuracy Metrics for Tanga-Voting Classifier

SN	Model Name	Performance (F1 Score)	Sensitivity	Balanced Accuracy
1.	Voting Classifier	0.32	1.00	0.50

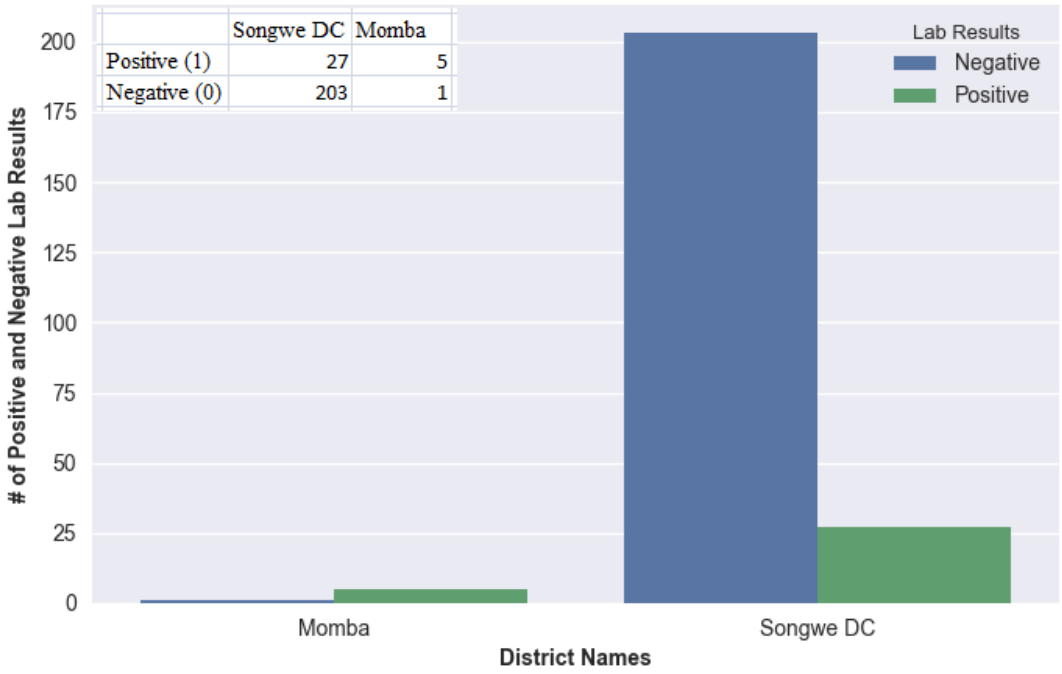


Figure 45: Patients Distribution as per LabResults (Positive and Negative) in Songwe Region

Table 23: Results of F1-Score, Sensitivity, Specificity and Balanced-Accuracy Metrics for Songwe-Voting Classifier

SN	Model Name	Performance (F1-Score)	Sensitivity	Balanced Accuracy
1.	Voting Classifier	0.25	0.995	0.50

4.4.3 Results of Model Assessment with its discussion

Figure 47 briefly presents results on how the 500 participants (whose profile was presented in Table 9) responded to the focus group discussion and interviewer-administered questionnaires during the model validation process. The assessment included all features proposed and implemented/ performed in the study from data collection to model development. The following briefly presents the results:

(i) Community Awareness on the Availability of ARML Prediction Model for Cholera Epidemics and HEMIS**:** 79% of the total participants were not aware of the existence of these systems in Tanzania. However, 90% of the participants after being given a brief knowledge on the systems agreed that they will assist in the prevention of cholera epidemics.

(ii) Knowledge on Cholera Transmission, Prevention and Control: 45% of the total participants, especially patients did not have proper knowledge on management of cholera disease in terms of its transmission, prevention and control.

(iii) Access to Safe Water and Sanitation Level: 65% of the total participants did not have a reliable source of safe water and lacked appropriate sewage and sanitation systems. Note the distribution of the 65% of the total population in terms of the rural and urban is at 8:5 or (40%:25%).

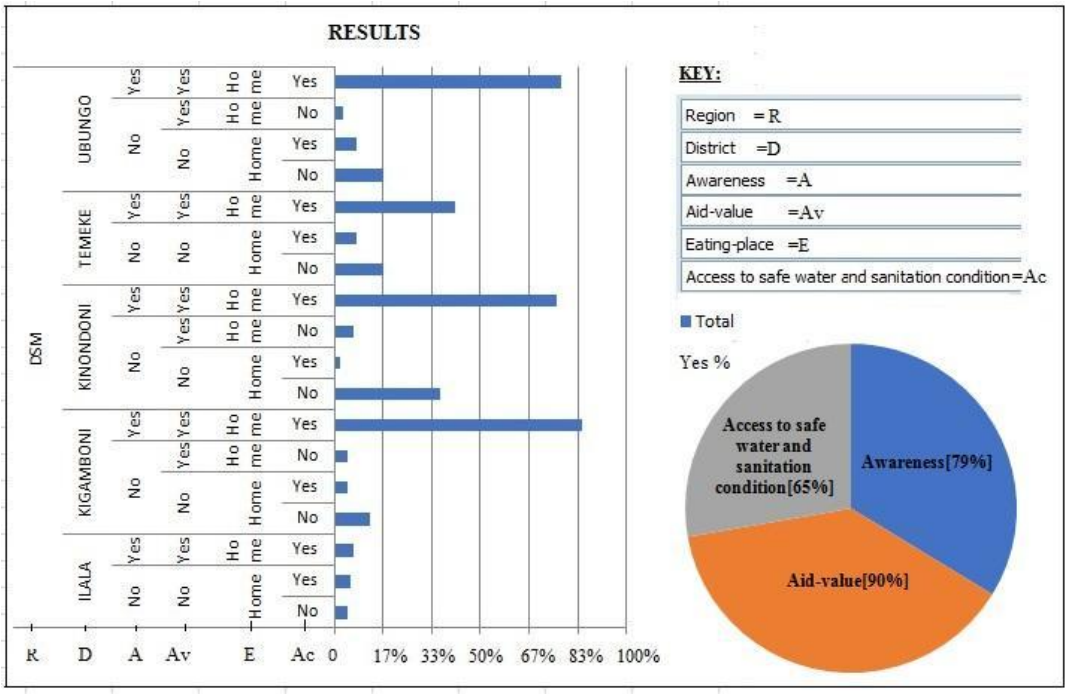


Figure 46: Response of 500 Participants

The results of assessment agree with several other findings related to e-Health and m-Health interventions in SSA such as (Igira *et al*., 2007; Karimuribo *et al*., 2011; Kwesigabo *et al*., 2012). However, the limited access to power supply, internet, ICT equipment such as computers, poor health-datasets and advanced knowledge on the use of ML were highlighted in the study result as challenges. The participants pointed the need to address these challenges

in order to ensure effective provision and access of ARML Prediction Model for cholera epidemics and HEMIS services in all hospitals in Tanzania.

4.4.4 Results of the developed Model Performance compared to Cholera Mathematical Models

Figure 49 briefly presents results on how the 20 participants (whose profile was presented in Table 10) responded to the focus group discussion and interviewer-administered questionnaires during the fourth model validation process. The following briefly presents the results:

(i) Usability: 80% of the total participants agreed that ML Models are better than Mathematical Model in terms of usability. This was so because, ML Models have wide community support, its codes are like English alphabets which can be easily used and applied by anybody and also, it's easy to teach yourself the codes and reuse the codes to apply to another datasets. Compared to Mathematical Models which requires the user to have knowledge on mathematic models and also, its model equations are not very easy to understand and use.

(ii) Extensibility: 90% of the total participants also, agreed that ML Models are better than Mathematical Model in terms of extensibility after observing the processes of adding an addition variable to each of the models. This was so because, the participants observed that in ML models you just add the variable into the dataset and the re-load the datasets and re-run the codes to get its model whereas for the Mathematical Model requires to write down another equation or add mathematically a variable into the existing equation with supporting parameters.

(iii) Computational Performance: The result of the third experiment for comparison of three cycles of Leo's Model has significantly shown the cost of additional variable on the model's prediction latency. As we can observe in Fig. 48 that the convergence of the three Leo's Model for infected population is not uniform, this is because the number of iterations in Z is more than in Y and the number of iterations in Y is more that in X. The one in blue represent the Z equations, in red represent the Y equations and the orange represent the X equations. Meaning the Model Z takes more prediction latency compared to Model Y and Model Y takes more prediction latency compared to Model X; Whereas for the ML Models has no

prediction latency variation, this is due to the fact that there are no a lot of changes in the models when two to four new variables are added.

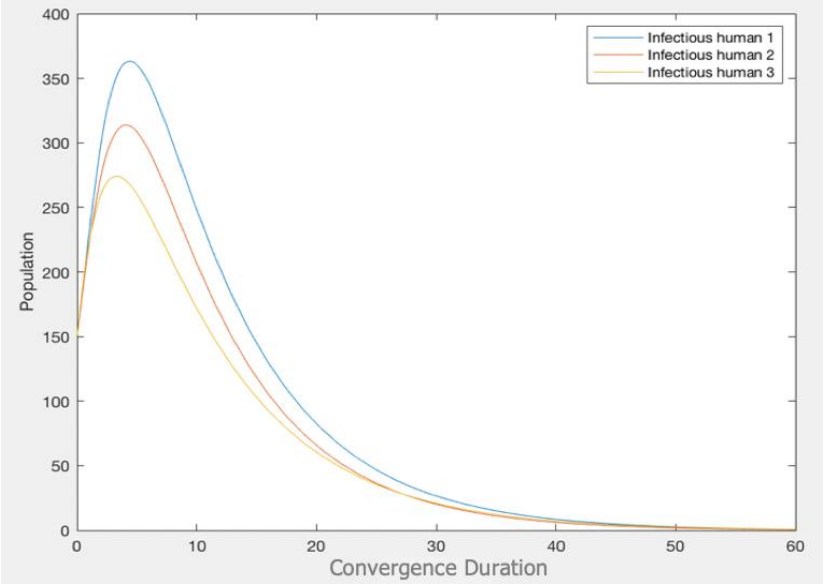


Figure 47: Convergence of Leo Models X, Y and Z

(iv) Accuracy: The participants agreed that both ML and Mathematical Models can generate accurate models at the rate of 50%:50% or at the ration of (1:1) respectively. This is due to the fact that each technique has various ways of optimising or improving the model's accuracy. It was also, noted and argued by the participants that, even though the mathematical modeling is losing momentum since the machine can handle large amounts of data in terms of storage and processing without the need to reduce them through the use of mathematical techniques (Kleinstreuer & Xu, 2016); researcher should still find a way on how to combine the two technologies to work together or complement each other.

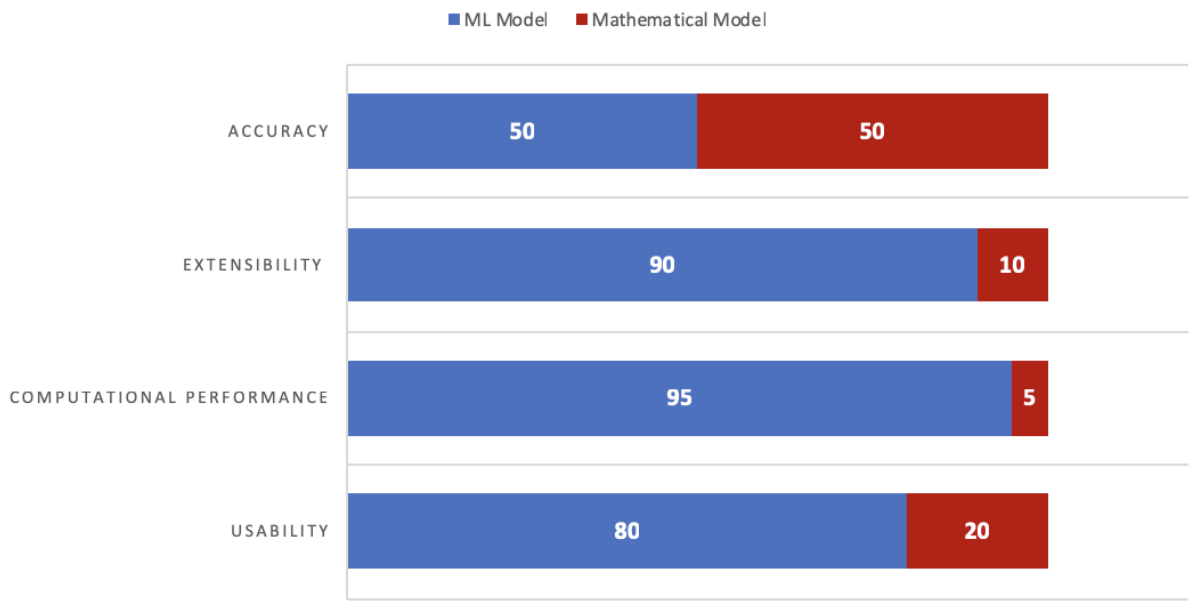


Figure 48: Response of the 20 Participants

4.5 Discussion

The study aimed at developing a reference ML model for prediction of cholera outbreaks based-on seasonal weather variables. The study was conducted in order to integrate seasonal weather variables into the prediction of cholera outbreaks. In the process, seasonal weather variables including rainfall, humidity and temperature were used with other variables as independent variables for cholera prediction. In addition, different ML techniques such as ensemble were used to manage the existing cholera dataset in Tanzanian health sector for effective prediction of an outbreak. Based on the analysis and prediction of the cholera dataset, its results indicated that seasonal temperature ranges from 22^{o}C to 32^{o}C, higher the rainfall level (>50 mm) and higher the humidity level (>75%) are the favourable conditions for cholera incidences/ outbreaks in Dar es Salaam. In addition, based on the analysis of cholera dataset, it was observed that cholera epidemic is affecting more in Kinondoni, Handeni, Korogwe, and Momba districts compared to other districts in the three regions i.e Dar es Salaam, Tanga, and Songwe regions. Furthermore, climatology-aware HIS was proposed and developed in order to assist proper collection of essential data for cholera prediction in the future; since the collected cholera dataset was of poor quality (missing values and variables). In addition, the climatology-aware HIS fulfils the intended goal of enhancing timely cholera epidemics analysis for prediction since, all essential required data such as; weather variables, geographical areas of patients' home and hospital, date and time

on the set of the patient can be collected effectively in a single platform. Furthermore, the system can screen the collected data in different tabular format (Microsoft excel format), which is useful for ML application. Therefore, the results of this study offer a number of practical implications to the health sector. Moreover, they also, give effective insights and assistance to cholera-patients, data collectors such as hospital receptionists (hospital receptionist is the one who feed information into the HMIS), medical officers, decision makers including policymakers, epidemiologists and other stakeholders dealing with environmental and waterborne related diseases/ epidemics. In brief, the following are some of important findings which led to the conduction of this research study:

(i) Tanzania is among the cholera endemic countries which experience frequent and recurring outbreaks; causing mortality rate of approximately 25 000 cases and 13 000 deaths yearly.

(ii) The existing health data in Tanzania is poorly collected hence most of its data have missing values and variables.

(iii) Health sectors in Tanzania and most of the developing countries are facing challenges in data management in terms of storage and data-usage for analysis and prediction.

(iv) Most of the hospitals in Tanzania are still using manual-based systems to conduct their daily activities.

(v) Most existing healthcare systems such as excel document and manual-based systems in Tanzania are not able to handle effective data analysis and prediction.

(vi) The health information system lacks inclusion or collection of all essential climatological which are necessary for effective analysis and prediction of cholera outbreaks in Tanzania.

(v) Machine learning techniques are very powerful, advanced and innovative tools for studying the dynamics of epidemics with a wide range of dynamic and variables such as data variability and inconsistency (Hepworth, Nefedov, Muchnik & Morgan, 2012).

The findings in the current study, such as the occurrence of cholera outbreaks are strongly associated or linked with seasonal temperature ranges from 22^{o}C to 32^{o}C, higher the rainfall level (>50 mm) and higher the humidity level (>75%) replicates the findings of a demonstrated relationship between cholera and climate conducted by (Constantin & Colwell, 2009) and (Siettos, 2018). This fact shows that our model produces similar and correct results interms of predicting an outbreak using seasonal weather variables in Tanzania. Similarly, this study replicates that the XGBoost classifier also performs well in predicting cholera outbreaks using weather variables such as temperature, rainfall and humidity as demonstrated in the study done by Siettos (2018) on Forecasting the 2017-2018 Yemen Cholera Outbreak. In addition, the current study replicates the fact that Random Forest and Decision Tree classifiers perform well when used in disease prediction, as demonstrated in the study done by Alotaibi (2019) on Implementation of Machine Learning Model to Predict Heart Failure Disease. Moreover, Ranjbar *et al.* (2011) have used the dimensions of water supply as predictors of cholera outbreak in the following studies; A Cholera Outbreak Associated with Drinking Contaminated Well Water and CHOLERA OUTBREAK: Assessing the outbreak response and improving preparedness, respectively. Certainly our study differs from other similar study done on Computational Prediction of Cholera Outbreak (Mubangizi *et al*., 2009) in Uganda hat its possible to timely predict cholera infection rates given history of infection in the area and climatic data. This is so, because Mubangizi *et al*. (2009) has enough and quality dataset from the Epidemiological and Surveillance Unit at the Uganda Ministry of Health and also the study was well funded by IBM Faculty Award which assisted in clearing any setbacks within the process.

Despite all measures being considered toward developing an effective model, the model cannot precisely predict a single value of weather variable for cholera prediction but rather a range of values. However, the findings of this study are novel due to the fact that the developed model is easily understandable, has less computational complexity, it can perform just as well and fast if integrated with the developed climatology-aware HIS. In addition, the model will easily predict the probability that there is occurrence of cholera or the patient has cholera due to the values of the weather variables compared to the use of laboratory test in the first diagnosis of cholera patients. In terms of less computational complexity is when you compare it with most of existing cholera prediction models in the country which are statistical and manual based. In addition, in terms of understandable is due to the ML codes used and this was verified by users as explained in Section 3.12 and Section 4.4.2 that ML codes are

easy to understand compared to mathematical theories that are applied in model development. Lastly, the model can perform just as well to give the required results based on the data used.

In addition, the results that we have previously described are sufficiently satisfactory in Tanzania, yet looking at nature and quality of the collected cholera dataset; the model has some difficulties to predict cholera outbreak based on specific time using the weather variables such as temperature, rainfall, humidity and wind; this is due to poor quality of dataset. Therefore, that study has focused on developing an effective model. Other limitations include; the study could not establish novel mechanism for handling imbalanced datasets, missing information, data inconsistency however we have innovatively combined and integrated the existing theories and strategies such as ensemble, in order to develop an effective reference ML model for prediction of cholera outbreak based on seasonal weather variables. In addition, the current study has used Dar es Salaam, Tanga and Songwe regions as study areas, which in country-size is a small scale to represent the whole of Tanzania.

Furthermore, the study shows that decision and policy makers should consider more into emphasizing proper health data collection with integration of all environmental variables including seasonal weather changes, inform the community especially in the rural areas on the mode of transmission, causes, recurrence of cholera epidemics in their areas, and interventions that need to be done towards its eradication. Also, encourage health and environmental sectors to work together towards the eradication of environmental related diseases such as cholera, promote the use and significances of ML technologies in the health and environmental sectors, and also, encourage cholera patients to visit hospital for treatments and cooperate well with giving out their personal and other essential details which will later assist in the analysis and prediction of diseases. In addition, the study recommends development of novel mechanisms for handling data challenges such as imbalance class and other uncertainties.

Moreover, the performance of the developed model against the existing cholera mathematical models showed that the more the number of variables in the cholera mathematical model the higher the running time of the model and when the results of several cholera mathematical models with different numbers of differential equations and or variables were plotted against their running time and as a result, they formed a linear graph as shown in Fig. 50. However, the study recommends that other researcher should look on how mathematical and ML model

theories can work together in order to combine their strength towards the development of an effective cholera prediction model.

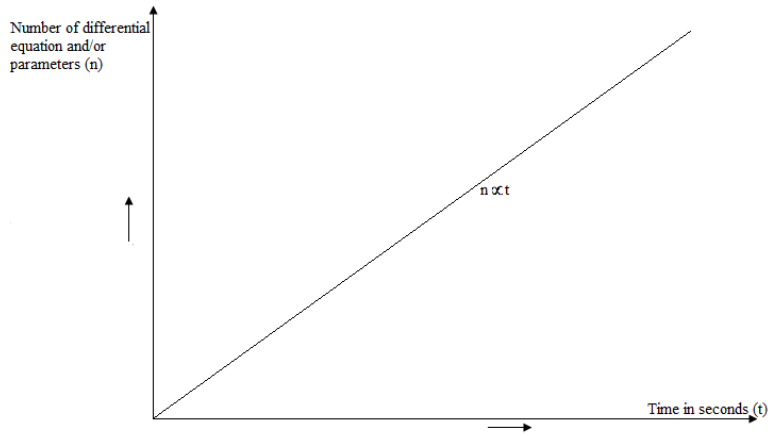


Figure 49: A Linear Graph Showing the Cost of Additional Variable on the Running Time

Lastly, it should be noted that effective prediction of cholera outbreaks through the use of ML is essential in order to limit the severity and duration of an outbreak. Hence, health sectors should collect timely all essential cholera predictors including seasonal weather variables in a single platform, which could prove useful in timely prediction of cholera epidemics in terms of size, duration, geographical location and its mortality impact in the area.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This chapter concludes what has been the focus of this study thus far, which is to develop an adaptive reference Machine Learning (ARML) model for prediction of cholera epidemics using seasonal weather changes in Tanzania. Through this development, the study aims at providing procedure (strategies and innovation applied) used, challenges faced and knowledge gained throughout the research study. Whereby, in the first objective, a brief review and description of required variables and also, the need and importance of using ML technologies in enhancing effective data analysis and prediction were provided.

The second specific objective was geared toward designing and development of the proposed model. The objective focused on demonstrating the ability of ML to analyse and predict cholera epidemics using environmental factors, particularly the seasonal weather changes in Tanzania. In addition, the procedure for data collection, pre-processing and also, dealing with imbalance datasets were presented. Lastly, it led into realization and development of the proposed model which can effectively predict the occurrence of cholera using seasonal weather variables in Tanzania.

The third specific objective aimed at validating the model's efficiency, robustness and impact to the selected study area and its stakeholders. Data from Tanga and Songwe regions were used for the validation process. The developed model in general attained an average rate of 52.68% robustness with regards to the used new set of datasets. In addition, study assessment showed that 90% of participants – patients and ICT, medical and environmental experts dealing with cholera variables – agreed that these interventions will add value in the prevention of cholera epidemics in Tanzania.

The global strategy for cholera management at the country level, has proposed to reduce the number of cases by 90% by 2030. One of the strategies suggested is the early detection and rapid response to contain cholera outbreaks since the existing disease surveillance systems such as Integrated Disease Surveillance and Response (IDSR), District Health Information System 2 (DHIS2) and traditional statistical techniques have enabled to manage epidemics such as cholera to some extent. However, when examining challenges of the collected cholera

datasets in Tanzania, such as imbalanced data, missing information, data inconsistency, multiple data-format and dynamic nature of its predictors such as weather variability. With the increase of number of epidemic predictors from the effect of climate change and the existence of limited number of the workforce in Tanzania's hospitals to handle manual-based and ineffective HIS for disease analysis and prediction; shows how difficult it was to formulate a suitable model. However, this study has managed to identify weather variables and other requirements that are important predictors for outbreak of cholera epidemics, develop climatology-aware HIS to facilitate timely collection of complete and quality datasets for cholera analysis and prediction, and also, develop an adaptive reference ML model for prediction of the occurrence of cholera epidemics based on weather variables. The developed model can be easily adapted and also, be used as a reference model by other data scientists and researchers. In addition, the study offers other useful insight of the nature of existing dataset in Tanzania and how such dataset should be treated towards mitigating the prediction performance of the models. Other contributions are as presented hereunder.

5.1.1 The Use of ML Model for Cholera Epidemics Analysis

The use of ML in this study has effectively analysed the dataset and managed to easily make users understand the data linkages and its seasonality. In addition, the analytical results produced in this study are capable of describing the data and its linkages; which are useful for cholera management. The obtained results of prediction indicate that temperature ranges from 22^{o}C to 32^{o}C, higher the rainfall level (>50 mm) and higher the humidity level (>75%) are the favourable conditions for cholera incidences. Therefore, through the use of this model with a given future weather variables; it is easy to predict the occurrence of cholera epidemics in an area.

However, during the data analysis the following limitations and challenges were observed which include:

(i) The in-ability of the collected data to develop time series prediction due to data inconsistency.

(ii) There is a low occurrence of cholera in Dar es Salaam based on the collected data. This fact could be attributed to the fact that reporting cholera has negative impact on business and other relationships such as trade, tourism etc. between countries.

(iii) Data analytical results were not able to detect patient's symptoms such as diarrhoea, wind directions and salary (income) as significant predictors of cholera incidences.

5.1.2 The Use of Climatology-aware HIS for Data Collection

The proposed and developed climatology-aware HIS can significantly assist collection of all essential data required in the analysis and prediction of cholera epidemics. The essential data includes variables such as; temperature, wind, humidity, rainfall, water (treated or untreated water), type of toilet used, level of education and cultural practices to mention a few. Therefore, since all required variables for cholera datasets will be collected and available in a single platform. Hence, timely analysis and prediction will be easily achieved, if the developed ML model will be integrated with the climatology-aware HIS or used data from the developed climatology-aware HIS.

5.1.3 Prediction of Cholera Epidemics Using ML Model

The procedure of developing prediction models has significantly assisted in understanding the existing challenges such as poor data collection in the health sector in Tanzania and also, learning of new opportunities that can be done in the health sector in order to overcome the existing challenges such as review of data collection systems in the health sector to mention a few. In addition, the prediction results can be used to make effective decisions, data-driven planning and policies, and also, assist in early awareness of occurrences of cholera based on the conditions of the weather. Moreover, the use of ML in this study has significantly shown that ML is capable of learning from a variety of observations, one by one without a need of making assumptions compared to the existing manual-based health systems and excel documents which were used for data analysis and prediction. Therefore, ML models can work effectively with data that have a wide number of variables, attributes and observations. Their capabilities have given assurance to the management and make effective use of existing data in the health sectors.

5.1.4 Data Screening for Decision Making Process

The developed climatology-aware HIS and model have enabled screening of collected data and results of cholera epidemic analysis and prediction processes. Therefore, through the screen of such data and their linkages; the epidemiologists, policy makers and all organs involved in cholera epidemics are able make proper decisions towards their activities such as

control measures, appropriate policies development for treatment and eradication of cholera epidemics.

5.1.5 Getting Insight from the Data and Realization of its Wide Range of Applicability

The analytical and prediction results of this study have given a wide understanding on the dynamics of Cholera variables, variations of weather changes and their linkages. In addition, the results have given insight towards understanding the extension on the use of the datasets into other applications such as the dynamics of treated and untreated water with cholera and other social economic factors. Furthermore, the collected data has assisted at large to come with various recommendations that need to be done by the Ministry of Health and Social Welfare, cholera-patients and e-Health NGOs to mention a few.

5.1.6 Uniqueness of Model Development Approach

The study has used a unique approach in developing the proposed model. This is because; according to the literature, there are several cholera predicting models that have been developed in the past in areas such as Yemen, where there is quality data, smooth seasonal weather changes and funds. Due to such favourable conditions, these models tend to be simpler. In the case of cholera epidemics in Tanzania, with poor dataset, limited funds, limited time and dynamics of cholera predictors such as variation of seasonal weather changes; necessitated the need of innovative strategies to enable the model to predict the outcomes effectively. With this regard, the study has used unique approach such as:

(i) Application of Adaptive Synthetic Sampling Approach (ADASYN) and Principal Component Analysis (PCA) in order to restore sampling balance and dimensions of the dataset.

(ii) The use of a wide number of algorithms in order to increase selection range.

(iii) The use of bagging classifier and ensemble method in order to enhance the robustness of the model.

(iv) The use of Friedman test for ranking and selecting the models or algorithms accurately.

5.2 Recommendations

In the process of addressing the specific objectives of this study in order to attain the main objective, a number of limitations arose. Some of these limitations include:

(i) Lack of unified platform to collect the required data.

(ii) A lot of missing information in the health-care data.

(iii) Un-proper and non- uniform format of weather and health-care data.

(iv) Lack of simultaneous collection of health-care data.

(v) Lack of enough health-care data.

(vi) Developed model and its formulation were based on the data collected.

Given the limitations, the study recommends a review of health-care systems in order to facilitate quality data collection through provision of awareness to all stakeholders in the data collection process which include patients and hospital officials; and also, strengthening of data collection tools. In addition, the study recommends collection of missing and upcoming data from both government and private hospitals in order to have enough data which could assist in comparing the results on extended datasets as well as formulating the challenge into time-series classification solution. Furthermore, the study recommends to the Ministry of Health and Social Welfare in Tanzania to not only form a special cholera management unit, but also, encourage the development and use of integrated or unified mechanisms for collection of both health-care and weather data in hospitals. Moreover, further studies should be done to compare the results on new sets of datasets from other countries. Lastly, further studies should look on how to secure ML models and also, implementation or integration of the satellite module as proposed in Fig. 32 and ARML model into the climatology-aware HIS.

Despite these limitations and recommendations, the achieved outcomes remain significant to the study area and the collected dataset. The study has added value to the body of knowledge of e-Health and ICT in dealing with poor quality of cholera datasets in least developing countries settings using ML techniques. It also has and will significantly manage the complexity of real-world problems such as data-driven analysis, decision making and

management of health-care datasets with various observations, prediction and eradication strategies of cholera epidemics at large scale in Tanzania. The study also, paves way for further analysis on the effects of seasonal weather variations on the dynamics of waterborne epidemics and other diseases through the use of ML techniques in least developing countries.

REFERENCE

Abuassba, A. O. M., Zhang, D., Luo, X., Shaheryar, A., & Ali, H. (2017). Improving classification performance through an advanced ensemble based heterogeneous extreme learning machines. *Computational Intelligence and Neuroscience, 8(2),* 1-20. https://doi.org/10.1155/2017/3405463

Africa, S. (2018). *Bulletin: Cholera and acute watery diarrhea outbreaks in eastern and southern Africa regional update for 2018.* Retrieved from https://reliefweb.int/report/united-republic-tanzania/bulletin-cholera-and-awd /united-republic-tanzania/bulletin-cholera-and-awd-outbreaks-eastern-and-southern-africa

Akosa, J. S. (2017). *Predictive accuracy: A misleading performance measure for highly imbalanced data: Special air service global forum, 942, 1–12*. Retrieved from https://www.semanticscholar.org/paper/Predictive-Accuracy-%3A-A-Misleading-Performance-for-Akosa/8eff162ba887b6ed3091d5b6aa1a89e23342cb5c

Alder, M. D., Aldo. A. M. (2001). *An introduction to mathematical modelling*. Retrieved from Heaven for Books. Com. https://doi.org/10.1002/bimj.19710130308

Ali, A., Qadir, J., Rasool, R. U., Sathiaseelan, A., Zwitter, A., & Crowcroft, J. (2016). Big data for development: Applications and techniques. *Big Data Analytics, 1*(1), 1-11. https://doi:10.1186/s41044-016-0002-4

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem. *International Journal of Advances in Soft Computing and Its Applications*, *7*(3), 176–204. https://doi.org/10.3387/cc2844

Andam, E. A., Apraku, L. O., Agyei, W., & Denteh, O. (2015). Modeling cholera dynamics with a control strategy in Ghana. *British Journal of Research, 2(1),* 030–041. https://doi.org/10.1234/cc5728

Azizi, M., & Azizi, F. (2010). History of cholera outbreaks in Iran during the 19th and 20th centuries. *Middle East Journal of Digestive Diseases,* 2(1), 51–5. https://doi.org/10.1181/cc2813

Azman, A. S., Rudolph, K. E., Cummings, D. A. T., & Lessler, J. (2013). The incubation period of cholera: A systematic review. *Journal of Infection*, *66*(5), 432–438. https://doi.org/10.1016/j.jinf.2012.11.013

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 10: Further nonparametric methods. *Critical Care*, *8*(3), 196–199. https://doi.org/10.1186/cc2857

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? *International Conference on Database Theory 1540*(99), *217-235.* https://doi.org/10.1007/3-540-49257-7_15

Bhavsar, H., & Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering*, *2(4),* 74–81. https://doi.org/10.1.1.492.6088

Biegel, B., & Kurose, J. F. (2016). *The national artificial intelligence research and development strategic plan (VIII-40).* Retrieved from https://www.nitrd.gov/news/National-AI-RD-Strategy-2019.aspx

Boisson, S., Engels, D., Gordon, B. A., Medlicott, K. O., Neira, M. P., Montresor, A., Velleman, Y. (2015). Water, sanitation and hygiene for accelerating and sustaining progress on neglected tropical diseases: A new global strategy 2015-20. *International Health*, *8*(Suppl 1), i19–i21. https://doi.org/10.1093/inthealth/ihv073

Boivin, G., Hardy, I., Tellier, G., & Maziade, J. (2000). Predicting influenza infections during epidemics with use of a clinical case definition. *Clinical Infectious Diseases*, *31*(5), 1166–1169. https://doi.org/10.1086/317425

Brauer, F. (2017). Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, *2*(2), 113–127. https://doi.org/10.1016/j.idm.2017.02.001

Breve, F. A., Ponti, M. P., & Mascarenhas, N. D. A. (2007). Multilayer perceptron classifier combination for identification of materials on noisy soil science multispectral images. *Proceedings of 2007 - 20th Brazilian Symposium on Computer Graphics and Image Processing, 1(1*), (239–244). https://doi.org/10.1109/SIBGRAPI.2007.10

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., & Rosenfeld, R. (2015). Flexible modeling of epidemics with an empirical bayes framework. *PLoS Computational Biology*, *11*(8), 1–18. https://doi.org/10.1371/journal.pcbi.1004382

Burnett, E., Dalipanda, T., Ogaoga, D., Gaiofa, J., Jilini, G., Halpin, A., Yen, C. (2016). Knowledge, attitudes and practices regarding diarrhea and cholera following an oral cholera vaccination campaign in the Solomon islands. *PLoS Neglected Tropical Diseases*, *10*(8), 1–9. https://doi.org/10.1371/journal.pntd.0004937

Bwire, G., Mwesawina, M., Baluku, Y., Kanyanda, S. S. E., & Orach, C. G. (2016). Cross-border cholera outbreaks in sub-Saharan Africa, the mystery behind the silent illness: What needs to be done? *PLoS One*, *11*(6), 1–15. https://doi.org/10.1371/journal.pone.0156674

Cain, J. W. (2017). Mathematical models in the sciences. *Molecular Life Sciences*, *8*(4) 1–6. https://doi.org/10.1007/978-1-4614-6436-5

Capasso, V., & Serio, G. (1978). A generalization of the Kermack-McKendrick deterministic epidemic model. *Mathematical Biosciences*, *42*(1–2), 43–61. https://doi.org/10.1016/0025-5564(78)90006-8

Castellazzo, A., Mauro, A., Volpe, C., & Venturino, E. (2012). Do demographic and disease structures affect the recurrence of epidemics? *Mathematical Modelling of Natural Phenomena*, *7*(3), 28–39. https://doi.org/10.1051/mmnp/20127303

Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International Journal of Environmental Research and Public Health, 15*(8), 1-88. https://doi.org/10.3390/ijerph15081596

Chan, C. H., Tuite, A. R., & Fisman, D. N. (2013). Historical epidemiology of the second cholera pandemic: Relevance to present day disease dynamics. *PLoS One*, *8*(8), 72497-72498. https://doi.org/10.1371/journal.pone.0072498

Chao, W. (2011). *Machine learning tutorial: Digital image signal processing*. Retrieved from http://disp.ee.ntu.edu.tw/~pujols/Machine Learning Tutorial

Codeço, C. T. (2001). Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir. *BioMedical Central Infectious Diseases*, *1*(1), 1-14. https://doi.org/10.1186/1471-2334-1-1

Constantin, G., Murtugudde, R., Sapiano, M. R. P., Nizam, A., Brown, C. W., Busalacchi, A. J., … Colwell, R. R. (2008). Environmental signatures associated with cholera epidemics. *Proceedings of the National Academy of Sciences*, *105*(46), 17676–17681. https://doi.org/10.1073/pnas.0809654105

Curran, K. G., Wells, E., Crowe, S. J., Narra, R., Oremo, J., Boru, W., … Kioko, J. (2018). Systems, supplies, and staff: A mixed-methods study of health care workers' experiences and health facility preparedness during a large national cholera outbreak, Kenya 2015. *BioMedical Central Public Health*, *18*(1), 1–12. https://doi.org/10.1186/ s12889-018-5584-5

Darcy, N., Elias, M., Swai, A., Danford, H., Rulagirwa, H., & Perera, S. (2015). eHealth strategy development: A case study in Tanzania. *Journal of Health Informatics in Africa*, *2*(2), 7(3). https://doi.org/10.12856/JHIA-2014-V2-I2-107

Dawson, P. M., & Perera, S. (1924). Mathematical and statistical analysis. *Magnesium Proceedings of the Royal Society of London Series B, Biological Sciences*, *282*(1808), 20150205-20150205. https://doi.org/10.1098/rspb.2015. 0205

Dawson, P. M., Werkman, M., Brooks-Pollock, E., & Tildesley, M. J. (2015). Epidemic predictions in an imperfect world: Modelling disease spread with partial data. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1808), 20150205–20150205. https://doi.org/10.1098/rspb.2015.0205

De Jongh, P. J. R., Larney, J., Mare, E., Van Vuuren, G. W., & Verster, T. (2017). A proposed best practice model validation framework for banks. *South African Journal of Economic and Management Sciences*, *20*(1), 1–15. https://doi.org/10.4102/ sajems. v20i1.1490

Demšar, J., Curk, T., Erjavec, A., Hočevar, T., Milutinovič, M., Možina, M., Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, *14(1)*, 1-2. https://doi.org/10.1098/rspb.2013.02

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, *185(7)*, 1–15. https://doi.org/10.1007/3-540-45014-9

Dietterich, T. G. (2009). Machine learning in ecosystem informatics and sustainability. *International Joint Conference on Artificial Intelligence, 5(3), 1-12*. https://doi.org/10.1007/978-3-540-75488-6_2

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the Association for Computing Machinery*, *55*(10), 77-78. https://doi.org/10.1145/2347736.2347755

Dorn, K. H., & Jobst, T. (2002). Innenreinigung von Rohrleitungssystemen aus Stahl. *Journal Fuer Oberflaechentechnik, 42*(5), 56–57. https://doi.org/10.1007/BF03251578

Du, K. L., & Swamy, M. N. S. (2014). Neural networks and statistical learning. *Quantum Machine Learning, 7452(3), 1-71*. https://doi.org/10.1007/978-1-4471-5571-3

Erik G. (2014). *Introduction to supervised learning, 1-5*. Retrieved from https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf%0Ahttp://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a

Escobar, L. E., Ryan, S. J., Stewart-Ibarra, A. M., Finkelstein, J. L., King, C. A., Qiao, H., & Polhemus, M. E. (2015). A global map of suitability for coastal *Vibrio cholerae* under current and future climate conditions. *Acta Tropical*, *149*, 202–211. https://doi.org/10.1016/j.actatropica.2015.05.028

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *09*(01), 1–16. https://doi.org/10.4236/jilsa.2017.91001

Fitzpatrick, C., & Engels, D. (2015). Leaving no one behind: A neglected tropical disease indicator and tracers for the sustainable development goals. *International Health*, *8*(Suppl 1), i15–i18. https://doi.org/10.1093/inthealth/ihw002

Foundation of National Science. (2007). *Discovery in complex or massive datasets: Common statistical themes, 1-23*. Retrieved from https://www.nsf.gov/mps/ dms/documents/ DiscoveryInComplexOrMassiveDatasets

Fung, I. C. H. (2014). Cholera transmission dynamic models for public health practitioners. *Emerging Themes in Epidemiology*, *11*(1), 1–11. https://doi.org/10.1186/1742-7622-11-1

Gawanmeh, A., Al-Hamadi, H., Al-Qutayri, M., Chin, S. K., & Saleem, K. (2016). Reliability analysis of healthcare information systems: State of the art and future directions. *2015 17th International Conference on E-Health Networking, Application and Services*, *8*(4), 68–74. https://doi.org/10.1109/HealthCom.2015.7454475

Glass, R. I., & Black, R. E. (2003). The epidemiology of cholera in India. *The Lancet*, *213*(5509), 677-678. https://doi.org/10.1016/s0140-6736(00)79162-8

Green, A. (2015). Violence in Burundi triggers refugee crisis. *The Lancet*, *386*(9994), 639–640. https://doi.org/10.1016/s0140-6736(15)61489-1

Hanson, V. (1996). *Introduction to the landmark thucydides: A comprehensive guide to the peloponnesian war*, *ix–xxxiii*. Retrieved from https://doi.org/10.1557/mrs.2011.276

Hashizume, M., Wagatsuma, Y., Faruque, A. S. G., Hayashi, T., Hunter, P. R., Armstrong, B., & Sack, D. A. (2008). Factors determining vulnerability to diarrhoea during and after severe floods in Bangladesh. *Journal of Water and Health*, *6*(3), 323–332. https://doi.org/10.2166/wh.2008.062

Hepworth, P. J., Nefedov, A. V., Muchnik, I. B., & Morgan, K. L. (2012). Broiler chickens can benefit from machine learning: Support vector machine analysis of observational epidemiological data. *Journal of the Royal Society Interface*, *9*(73), 1934–1942. https://doi.org/10.1098/rsif.2011.0852

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced datasets: From sampling to classifiers. *Wiley Online Library*, 9(3) 43–59. https://doi.org/10.1002/9781118646106.ch3

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, *61*(1), 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Hulse, V. J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. *Proceedings of the 24th international conference on Machine learning, 7*(3), 1-14. https://doi:10.1145/1273496.1273614

Huppert, A., & Katriel, G. (2013). Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection*, *19*(11), 999–1005. https://doi.org/10. 1111/ 1469-0691.12308

Igira, F. T., Titlestad, O. H., Lungo, J. H., Makungu, A., Khamis, M. M., Sheikh, Y., Braa, J. (2007). Designing and implementing hospital management information systems in developing countries: Case studies from Tanzania-Zanzibar. *Health Informatics in Africa*, *9*(2), 1-5. https://doi:10.1145/1273496.1273684

Impouma, B., Kasolo, F., Yada, A., Yoti, Z., Yaya, S., Woodfill, C., & Roungou, J. B. (2007). Recurring epidemics in the WHO African region: Situation analysis, preparedness and response. *The Africa Health Monitor and Disease Control*, *7*(15), 42–47. https://doi.org/10.2366/cc2877

Joachim, R., & Karl, E. K. (2002). *Vibrio cholerae* and cholera: Out of the water and into the host. *Fems Microbiology Reviews*, *26*(2), 125–139. https://doi.org/10.7726/science.a444

Jacobi, P., Amend, J., & Kiango, S. (2000). *Urban agriculture in Dar es Salaam: Providing an indispensable part of the diet*, *257–283*. Retrieved from https://agris.fao.org/agris-search/search.do?recordID=NL2001003063

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 1-14. https://doi.org/10.1126/science.118415

Sarmanova, A., & Albayrak, S. (2013). Alleviating class of imbalance problem in data mining. *2013 21st Signal Processing and Communications Applications Conference, 8*(3), 1–17. https://doi:10.1109/siu.2013.6531574

Jutla, A., Whitcombe, E., Hasan, N., Haley, B., Akanda, A., Huq, A., … Colwell, R. (2013). Environmental factors influencing epidemic cholera. *The American Journal of Tropical Medicine and Hygiene*, *89*(3), 597–607. https://doi.org/10. 4269/ajtmh.12-0721

Kajirunga, A., & Kalegele, K. (2015). Analysis of activities and operations in the current e-health landscape in Tanzania: Focus on interoperability and collaboration. *International Journal of Computer Science and Information Security*, *13*(6), 49–54. https://doi:10.1109/siu.2013.6531574

Karaolis, D. K. R., Lan, R., & Reeves, P. R. (1995). The sixth and seventh cholera pandemics are due to independent clones separately derived from environmental, nontoxigenic, non-O1 *Vibrio cholerae*. *Journal of Bacteriology*, *177*(11), 3191–3198. https://doi.org/10.1128/ jb.177.11.3191-3198.1995

Karimuribo, E. D., Mboera, L. E. G., Mbugi, E., Simba, A., Kivaria, F. M., Mmbuji, P., & Rweyemamu, M. M. (2011). Are we prepared for emerging and re-emerging diseases? Experience and lessons from epidemics that occurred in Tanzania during the last five decades. *Tanzania Journal of Health Research*, *13*(5 SUPPL.ISS), 1–14. https://doi.org/10. 4314/thrb.v13i5.8

Kay, M., Patel, S. N., & Kientz, J. A. (2015). How good is 85%?: A survey tool to connect classifier evaluation to acceptability of accuracy. *Proceedings of the Association Computing Machinery Conference on Human Factors in Computing Systems, 5(1),* 347–356. https:// doi.org/10.1214/16-101S908

Kelly, J. E. (2015). Computing, cognition and the future of knowing. *Proceedings of the IBM on White Paper, 7*(1), 1-12. https://doi.org/10.1016/j.computeplace.2015.04.006

Koepke, A. A., Longini, I. M., Halloran, M. E., Wakefield, J., & Minin, V. N. (2016). Predictive modeling of cholera outbreaks in Bangladesh. *Annals of Applied Statistics, 10*(2), 575–595. https://doi.org/10.1214/16-AOAS908

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31(1),* 249–268. https://doi.org/10.1115/1.1559160

Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R. R., & Buyya, R. (2016). The anatomy of big data computing. *Software - Practice and Experience*, *46*(1), 79–105. https://doi.org/ 10.1002/spe.2374

Kwesigabo, G., Mwangu, M. A., Kakoko, D. C., Warriner, I., Mkony, C. A., Killewo, J., … Freeman, P. (2012). Tanzania's health system and workforce crisis. *Journal of Public Health Policy*, *33*(SUPPL.1), 35–45. https://doi.org/10.1057/1.2012.55

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care and Pain*, *8*(6), 221–223. https://doi.org/10.1093/ bjaceaccp/mkn041

Lan, R., & Reeves, P. R. (2006). Pandemic spread of cholera: Genetic diversity and relationships within the seventh pandemic. *Society*, *40*(1), 172–181. https://doi.org/10.1128/JCM.40.1.172

Leckebusch, G. C., & Abdussalam, A. F. (2015). Health and Place Climate and socioeconomic influences on interannual variability of cholera in Nigeria. *Health & Place*, *34*, 107–117. https://doi.org/10.1016/j.healthplace.2015.04.006

Ledder, G. (2013). Dynamics of single populations. *Mathematics for the Life Sciences*, *7*(2), 1–10. https://doi.org/10.1007/978-1-4614-7276-6

Lent, C. S. (2013). Learning to program with matlab building graphical user interface tools. *John Wiley and Sons Inc.*, 1(1), 309-310. https://doi.org/10.1007/s13398-014-0173-7.2

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets. *Cambridge University Press*, *9*(3), 1–12. https://doi.org/10. 1017/CBO9781139924801

Lessler, J., Moore, S. M., Luquero, F. J., McKay, H. S., Grais, R., Henkens, M., … Azman, A. S. (2018). Mapping the burden of cholera in sub-Saharan Africa and implications for control: An analysis of data across geographical scales. *The Lancet*, *391*(10133), 1908–1915. https://doi.org/10.1016/S0140-6736(17)33050-7

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/ j.jesp.2013. 03.013

Li, R., Wang, W., & Di, Z. (2017). Effects of human dynamics on epidemic spreading in Côte d'Ivoire. *Physica A: Statistical Mechanics and Its Applications*, *467*, 30–40. https://doi.org/ 10. 1016/j.physa.2016.09.059

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311. https://doi.org/10.1016/j.eswa.2012.02.063

Lipp, E. K., Huq, A., & Colwell, R. R. (2002). Effects of global climate on infectious disease: The cholera model effects of global climate on infectious disease. *Clinical Microbiology Reviews*, *15*(4), 757–770. https://doi.org/10. 1128/CMR.15.4.757

Lloyd, M. (2005). Towards a definition of the integration of information and communication technology in the classroom. *Proceedings Australian Association for Research in Education '05 Education Research - Creative Dissent: Constructive Solutions*, 1(1), 1–18. https://doi. org/ 10.1063/1.2130520

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv*, *9*(4), 1–10. https://doi.org/10.13140/2.1.1570.5928

Lugomela, C., Lyimo, T. J., Namkinga, L. A., & Moyo, S. (2015). Co-variation of cholera with climatic and environmental parameters in coastal regions of Tanzania. *Western Indian Ocean Marine Science*, *13*(1), 93–105. https://doi.org/10.1016/S1995-7645(11)60149-1

Lund, A. J., Keys, H. M., Leventhal, S., Foster, J. W., & Freeman, M. C. (2015). Prevalence of cholera risk factors between migrant Haitians and Dominicans in the Dominican Republic. *Revista Panamericana de Salud Publica*, *37*(3), 125–132. https://doi.org/ https://www.ncbi. nlm.nih.gov/pubmed/25988248

Mandal, S., Mandal, M. D., & Pal, N. K. (2011). Cholera: A great global concern. *Asian Pacific Journal of Tropical Medicine*, *4*(7), 573–580. https://doi.org/10.1016/S1995-7645(11)60149-1

Manzi, F., Schellenberg, J. A., Hutton, G., Wyss, K., Mbuya, C., Shirima, K., … Schellenberg, D. (2012). Human resources for health care delivery in Tanzania: A multifaceted problem. *Human Resources for Health*, *10*(1), 1-3. https://doi.org/10.1186/1478-4491-10-3

Mari, L., Bertuzzo, E., Righetto, L., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I., & Rinaldo, A. (2012). Modelling cholera epidemics: The role of waterways, human mobility and sanitation. *Journal of the Royal Society Interface*, *9*(67), 376–388. https://doi.org/10.1098/ rsif.2011.0304

Mrosig, F., Huber, N., & Kounev, S. (2014). Architecture-level software performance models for online performance prediction. *Science of Computer Programming, 90(1),* 71-92. doi:10.1016/j.scico.2013.06.004

Medsker, B., Forno, E., Simhan, H., Juan, C., & Sciences, R. (2016). Prenatal analysis. *Health Human Service Public Access, 70(12), 773–779.* https://doi.org/10.1097/OGX.0000000000000256

Mghamba, J., Mboera, L., Krekamoo, W., Senkoro, K., Rumisha, S., Shayo, E., & Mmbuji, P. (2011). Challenges of implementing an integrated disease surveillance and response strategy using the current health management information system in Tanzania. *Tanzania Journal of Health Research*, *6*(2), 57–63. https://doi.org/10.4314/thrb.v6i2. 14243

Midani, F. S., Weil, A. A., Chowdhury, F., Begum, Y. A., Khan, A. I., Debela, M. D., … Larocque, R. C. (2018). Human gut microbiota predicts susceptibility to *Vibrio cholerae* infection. *Journal of Infectious Diseases*, *218*(4), 645–653. https://doi.org/10.1093/infdis/ jiy192

Miller, C. J., Feachem, R. G., & Drasar, B. S. (1985). Cholera epidemiology in developed and developing countries: New thoughts on transmission, seasonality, and control. *The Lancet*, *325*(8423), 261–263. https://doi.org/10.1016/S0140-6736(85)91036-0

Miller, D. D., & Brown, E. W. (2017). Artificial intelligence in medical practice: The question to the answer? *American Journal of Medicine*, *131*(2), 129–133. https://doi.org/10.1016/j. amjmed.2017.10.035

Miller, R., Acton, C., Fullerton, D. A., & Maltby, J. (2007). *Stastistical package for the social scientists Book*, 1–353. Retrieved from https://gtu.ge/Agro-Lib/1%20

Mitchell, R., & Frank, E. (2017). Accelerating the xgboost algorithm using graphics processing unit computing. *Peer-reviewed Journal of Computer Science, 7(3), 126-127.* https://doi.org/10.7717/peerj-cs.127

Morof, D., Cookson, S. T., Laver, S., Chirundu, D., Desai, S., Mathenge, P., … Handzel, T. (2013). Community mortality from cholera: Urban and rural districts in Zimbabwe. *American Journal of Tropical Medicine and Hygiene*, *88*(4), 645–650. https://doi.org/10.4269/ajtmh.11-0696

Mugerezi, E. (2000). An environmental management information system for Iringa municipality, Tanzania implementation challenges. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XXXIV*, 78–86. Retrieved from https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi

Mukhopadhyay, A. (2015). Mapping of cholera cases using satellite-based recording systems to investigate the outbreak. *Indian Journal of Medical Research*, *142*(5), 508-509. https://doi.org /10 .4103/0971-5916.171269

Naha, A., Chowdhury, G., Ghosh-Banerjee, J., Senoh, M., Takahashi, T., Ley, B., … Mukhopadhyay, A. K. (2013). Molecular characterization of high-level-cholera-toxin-producing el tor variant *Vibrio Cholerae* strains in the Zanzibar archipelago of Tanzania. *Journal of Clinical Microbiology*, *51*(3), 1040–1045. https://doi.org/10.1128/JCM.03162-12

Narra, R., Maeda, J. M., Temba, H., Mghamba, J., Nyanga, A., Greiner, A. L., … Quick, R. E. (2017). Ongoing cholera epidemic - Tanzania, 2015-2016. *Morbidity and Mortality Weekly Report*, *66*(6), 177–178. Retrieved from https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://search.proquest.com/docview/1870842311?accountid=15115%0Ahttp://vr2pk9sx9w.search.serialssolutions.com?ctx_ver=Z39.88-2004&ctxenc=info:ofi /enc:UTF-8&rfr_id=info:sid/ProQ%3Anahs&rft_val_fmt=info:ofi/f

Nicholas Cuneo, C., Sollom, R., & Beyrer, C. (2017). The cholera epidemic in Zimbabwe, 2008–2009: A review and critique of the evidence. *Health and Human Rights*, *19*(2), 249–264. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5739374

Novais, R. C., Coelho, A., Salles, C. A., & Vicente, A. C. P. (1999). Toxin-co-regulated pilus cluster in non-O1, non-toxigenic *Vibrio cholerae*: Evidence of a third allele of pilin gene. *Fems Microbiology Letters*, *171*(1), 49–55. https://doi.org/10.1016/S0378-1097(98)00574-6 October, 2017. (2018), (18)

Olivera, A. R., Roesler, V., Iochpe, C., Schmidt, M. I., Vigo, Á., Barreto, S. M., & Duncan, B. B. (2017). Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes -Brasil: Accuracy study. *Sao Paulo Medical Journal*, *135*(3), 234–246. https://doi.org/10.1590/1516-3180.2016.0309010217

Peng, R. D. (2015). R programming for data science. *The R Project and Foundation*, *7*(5), 130-132. https://doi.org/10.1073/pnas.0703993104

Pezeshki, Z., Tafazzoli-Shadpour, M., Nejadgholi, I., Mansourian, A., & Rahbar, M. (2016). Model of cholera forecasting using artificial neural network in Chabahar city, Iran. *International Journal of Enteric Pathogens*, *4*(1), 40-45. https://doi.org/10.17795/ijep31445

Pompe, P. P. M., & Feelders, A. J. (1997). Using machine learning, neural networks, and statistics to predict corporate bankruptcy. *Computer-Aided Civil and Infrastructure Engineering*, *12*(4), 267–276. https://doi.org/10.1111/0885-9507.00062

Quarteroni, A. (2009). Mathematical models in science and engineering. *Notices of the American Meteorological Society*, *56*(1), 10–19. Retrieved from http://www.ams.org/notices/200901/tx090100010p

Raghupathi, V., & Raghupathi, W. (2017). Preventive healthcare: A neural network analysis of behavioral habits and chronic diseases. *Healthcare*, *5*(1), 1-8. https://doi.org/10. 3390/ healthcare5010008

Rebaudet, S., Sudre, B., Faucher, B., & Piarroux, R. (2013). Environmental determinants of cholera outbreaks in inland Africa: A systematic review of main transmission foci and propagation routes. *Journal of Infectious Diseases*, *208*(SUPPL. 1), 1-15. https://doi.org/10. 1093/ infdis/jit195

Richardson, B. (1979). Limitations on the use of mathematical models in transportation policy analysis. *Motor Vehicle Manufacturers Association*, *8*(3) 1–13. Retrieved from http://deepblue. lib.umich.edu/handle/2027.42/509

Saif-Ur-Rahman, K. M., Parvin, T., Bhuyian, S. I., Zohura, F., Begum, F., Rashid, M. U., … George, C. M. (2016). Promotion of cholera awareness among households of cholera patients: A randomized controlled trial of the cholera-hospital-based-intervention-for-7 days (chobi7) intervention. *American Journal of Tropical Medicine and Hygiene*, *95*(6), 1292–1298. https://doi.org/10.4269/ajtmh.16-0378

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38. https://doi.org/10.14569/IJARAI.2013.020206

Schaer, R., Müller, H., & Depeursinge, A. (2016). Optimized distributed hyperparameter search and simulation for lung texture classification in CT using Hadoop. *Journal of Imaging*, *2*(2), 18-19. https://doi.org/10.3390/jimaging2020019

Schroeder, R., Meyer, E., & Taylor, L. (2013). Big data and the uses and disadvantages of scientificity for social research. *Policy & Internet,* 6(4), 418-444. doi:10.1002/1944-2866.poi378

Health Security and Emergencies Cluster Outbreaks and Emergencies. (2016). Overview of reported public health and environmental services in World Health Organisation African region. *World Health Organisation*, *6*(4). Retrieved from http://www.afro.who.int/pt/grupos-organicos-e-programas/ddc/alerta-e-resposta-epidemias-e-pandemias/outbreak-news/5005-outbreaks-and-emergencies-in-the-who-african-region-bulletin-vol-6-issue-4-30-june-2016.html

Seleman, A., & Bhat, M. G. (2016). Multi-criteria assessment of sanitation technologies in rural Tanzania: Implications for program implementation, health and socio-economic improvements. *Technology in Society*, *46*, 70–79. https://doi.org/10.1016/j.techsoc.2016. 04. 003

Sharma, S., & Agrawal, J. (2013). Classification through machine learning technique: C4.5 algorithm based on various entropies. *International Journal of Computer Applications*, *82*(16), 975–8887. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download? doi=10. 1.1.402.2386&rep=rep1&type=pdf

Sharma, T., & Verma, S. (2017). Intelligent heart disease prediction system using machine learning: A review, *4*(2), 94–97. Retrieved from https://www.semanticscholar.org

Shemsanga, C., Nyatichi, A., & Gu, Y. (2010). The cost of climate change in Tanzania: Impacts and adaptations. *American Science*, *6*(3), 182–196. Retrieved from http://www.jofamericanscience.org/journals/am-sci/am0603/24_2189_climate_am0603_182_196

Shuai, Z., & van den Driessche, P. (2015). Modelling and control of cholera on networks with a common water source. *Journal of Biological Dynamics*, *9*, 90–103. https://doi.org/10.1080/17513758.2014.944226

Siettos, C. I., & Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence*, *4*(4), 295–306. https://doi.org/10.4161/viru.24041

Singh, D. (2013). A study on the use of non-parametric tests for experimentation with cluster analysis. *International Journal of Engineering and Management Research*, *6*(6), 64–72. https://doi.org/10. 1002/j.1681-4835. 2008.tb00227.x

Sinha, S., Chakraborty, R., De, K., Khan, A., Datta, S., Ramamurthy, T., … Nair, G. B. (2002). Escalating association of *Vibrio cholerae* O139 with cholera outbreaks in India. *Journal of Clinical Microbiology*, *40*(7), 2635–2637. https://doi.org/10.1128/JCM.40.7.2635-2637.2002

Smith, M., Madon, S., Anifalaje, A., Lazarro-Malecela, M., & Michael, E. (2008). Integrated health information systems in Tanzania: Experience and challenges. *The Electronic Journal of Information Systems in Developing Countries*, *33*(1), 1–21. https://doi.org/10. 1002/j.1681-4835. 2008.tb00227.x

Sonke, J., Rollins, J., Brandman, R., & Graham-Pole, J. (2009). The state of the arts in healthcare in the United States. *Arts & Health*, *1*(2), 107–135. https://doi.org/10.1080/17533010903031580

Soto, S. M. (2009). Human migration and infectious diseases. *Clinical Microbiology and Infection*, *15*(SUPPL. 1), 26–28. https://doi.org/10.1111/j.1469-0691.2008.02694.x

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631–643. https://doi.org/10.1093/bioinformatics/bti033

Stephenson, J. (2009). Cholera in Zimbabwe. *The Journal of the American Medical Association*, *301*(3), 263–263. https://doi.org/10.1001/jama.2008.961

Sustainable Accounting Standards Board. (2013). *Conceptual framework of the sustainability accounting standards board*. Retrieved from sasb.org

Sutton, C. D. (2005). Conceptual framework of the sustainability accounting standards board. *International Journal of Engineering and Management Research, 7161*(04), 1–27. Retrieved from https://doi.org/10.1016/S0169-7161(04)24011-1

Taylor, B., Chinamo, E., (2009). *Extended analysis of women, children and the water, sanitation and hygiene sector in Tanzania*. Retrieved from http://www.tzdpg.or.tz/fileadmin/documents/ dpg_ internal/dpg_working_groups_ clusters/cluster_2/education/3-Core_Documents/2.03-Joint_ Education_Sector_ Review /JESR09_MKUKUTA_SITAN09_WASH

Thessen, A. (2016). Adoption of machine learning techniques in ecology and earth science. *One Ecosystem*, *1*, e8621. https://doi.org/10.3897/oneeco.1.e8621

Thomson, J., O'Boyle, M., Fursin, G., & Franke, B. (2010). Reducing training time in a one-shot machine learning-based compiler. *International Journal of Artificial Intelligence in Bioinformatics*, *5898 LNCS*, 399–407. https://doi.org/10.1007/978-3-642-13374-9_28

Trærup, S. L. M., Ortiz, R. A., & Markandya, A. (2010). The health impacts of climate change: a study of cholera in Tanzania. *Basque Centre for Climate Change*, 7(12), 43–47. https://doi.org/10.3390/ ijerph8124386

Trærup, S. L. M., Ortiz, R. A., & Markandya, A. (2011). The costs of climate change: A study of cholera in Tanzania. *International Journal of Environmental Research and Public Health*, *8*(12), 4386–4405. https://doi.org/10.3399/ijerph8124512

Walldorf, J. A., Date, K. A., Sreenivasan, N., Harris, J. B., & Hyde, T. B. (2017). Lessons learned from emergency response vaccination efforts for cholera, typhoid, yellow fever, and ebola. *Emerging Infectious Diseases*, *23*(December), S210–S216. https://doi.org/10.3201/ eid2313.170550

Weill, F., Domman, D., Njamkepo, E., Tarr, C., Rauzier, J., Fawal, N., … Dougan, G. (2017). pandemic of cholera in Africa. *Science*, *789* (November), 785–789. Retrieved from http://www.jmlr.org/papers/v19/17-285.html

Were, V., Person, B., Kola, S., Nygren, B., Ayers, T., Date, K., & Quick, R. (2013). Evaluation of a rapid cholera response activity--nyanza province, Kenya, 2008. *Journal of Infectious Diseases*, *208*(suppl 1), S62–S68. https://doi.org/10.1093/ infdis/ jit198

Werner, K., Brandon, D., Clark, M., & Gangopadhyay, S. (2005). Incorporating medium-range numerical weather model output into the ensemble streamflow prediction system of the national weather service. *Journal of Hydrometeorology*, *6*(2), 101–114. https://doi.org/10. 1175/jhm411.1

APPENDICES

This section consists of appendices and questionnaire that were used in data collection and system testing and evaluation. While Appendix 1 shows the data collection photos, Appendix 2 shows the system testing and evaluation photos, Appendix 3 the questionnaire, Appendix 4 the mobile application interfaces of the developed system, and Appendix 5 shows the GIS based HIS interfaces of the developed system and questionnaire demonstrates the type of question that were given to users for system testing and evaluation.

Appendix 1: Data Collection

Before developing the proposed system, the problem domain that assisted us in designing and implementing the developed model was thoroughly studied. The photos below were taken during preliminary study stages.

Figure 50: Talking to a Medical Expert and one of the Patients during Data Collection

Appendix 2: Model Testing and Evaluation

After model development and implementation, model testing and evaluation was performed as a step towards fulfilling the second objective of the study.

Figure 51: Performing Installation of the Model into the Server

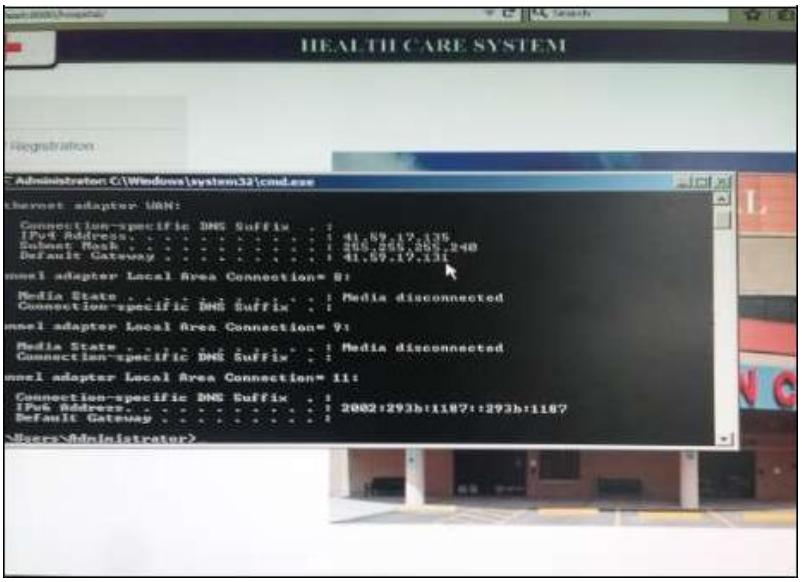


Figure 52: Some of the Configuration Codes into the Control Panel

Appendix 3: Questionnaires

I am Judith Leo, a Ph.D. candidate from Nelson Mandela African Institution of Science and Technology (NM-AIST) - Arusha. I am currently doing a research on developing ARPM. This questionnaire is aimed at assessing the perspectives of health-care and environmental workers as well as cholera and diarrhoea patients on its feasibility, user-acceptance, performance, complexity, impact and awareness of using an integrated weather variables-based healthcare model to enhance timely cholera epidemics analysis and prediction in Tanzania. The following are some of the sample questions.

Questionnaire Number:

Date:

Name of the working location:

A. Social-Economic Profile of the Respondent

1. Full name of the respondent: …….
2. Gender *(Mark only one)*

   Male

   Female

3. Age *(Mark only one)*

   Below 18 years

   18 up to 50 years

   Above 50 years

4. Educational Qualification: *(Choose number from table below)*

Below Certificate=1	Certificate=2	Diploma=3	Degree=4	Above degree=5

5. Professional:
6. Do you own Smartphone? *(Mark only one)*

   Yes

   No

7. Have you ever connected your phone to any internet bundle? *(Mark only one)*

   Yes

   No

8. Have you ever installed any app in your smartphone? *(Mark only one)*

   Yes

   No

9. What type of information do you obtain through your mobile phones? (Check all that apply)

   Health

   Sports

   Politics

   Fashion

   Agriculture

   I do not receive any information

   Others ……

10. What is your opinion about receiving health information via mobile phone and healthcare models/systems? *(Mark only one and add comment if any).*

    It is very important

    It is not important

    I do not know

    Add comment if any…….
    ……
    ……
    ……

11. What type of health and environmental information related to cholera epidemics would you want to send and/or receive using your mobile phones and healthcare model? *(Please briefly explain)* …….

    ……
    ……
    ……

……
……

12. What measures do you take to secure your confidential/ personal information received/ send in healthcare systems/ models? *(Please briefly explain)*
……
……
……
……
……

13. Do you think that, this initiative (ARML and HEMIS intervention) will improve data analysis and reduce cholera epidemics in Tanzania? *(Mark only one)*

    Yes …… if Yes explain how……...

    No …… if Yes explain why……...

14. Do you think that this initiative will reduce errors in capturing patient data? *(Mark only one)*

    Yes

    No

15. Do you think that this initiative will assist in enhancing cholera data analysis and prediction in Tanzania? *(Mark only one)*

    Yes

    No

16. Are the ML codes used easy to understand compared to mathematical codes used in the development of cholera model? Here only non- ICT and non-mathematical participants were tested, in terms of mathematical model Codeco's model equations were used *(Mark only one)*

    Yes

    No

B. Information about the initiative (Please answer questions that concerns you as regards to expert levels)

In this section of questionnaire, there will be technical questions to be answered by normal users, ICT experts and medical experts). The questions to be answered by normal users will be indicated by *, ICT experts **, Medical Experts *** and for all users ****.

Please indicate the level of agreement by ticking in the box, corresponding to the row and column					
	Strongly Agree	Agree	Neither Agree nor disagree	Disagree	Strongly Disagree
I found it easy to learn and operate this model ****					
The interface of the framework of this model is interactive ****					
I think that, I would like to continue using this model ****					
I think technical support is needed to operate this model ****					
I would imagine that most users will learn quickly to operate this model ****					
I think the model will assist in timely data analysis and presentation** and ***					
I found various functions of the model were well integrated ** and ***					
I found too much inconsistencies in this model ** and ***					
I would be happy to use the model for health services such as cholera analysis and prediction ****					
I think the mode will be useful to me ****					

Appendix 4: Sample Codes Used During the Research Study

Source codes from classifiers.py file

```
#Classifiers
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn import svm
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import ExtraTreesClassifier
from xgboost import XGBClassifier
from sklearn.neural_network import MLPClassifier

# Apply oversampling and scaling
from imblearn.over_sampling import ADASYN
X_resampled, y_resampled = ADASYN().fit_resample(X, y)
(y_resampled==1).sum()/len(y_resampled) # more balanced
min_max_scaler = preprocessing.MinMaxScaler()
X_resampled = min_max_scaler.fit_transform(X_resampled)

# Testing Different Scenarios: with/without oversampling, with/without PCA and combined
def classifiers(
        X, y, metrics=[sensitivity_score, specificity_score, balanced_accuracy_score],
        oversampling=False, use_pca=False, pca_components=3, pca_kernel="linear", cv_splits=10):
    # I commented out classifiers that underformed by a lot
    models = [
        (XGBClassifier, {"n_jobs":-1, "eta": .2, "scale_pos_weight":1.5}),
        (KNeighborsClassifier,{"n_neighbors": 9,}),
#         (SVC, {"gamma":"scale", "C": 5}),
#         (GaussianNB, {}),
#         (LogisticRegression, {"solver":"saga"}),
        (tree.DecisionTreeClassifier, {}),
        (RandomForestClassifier, {"n_estimators": 100}),
        (ExtraTreesClassifier, {"n_estimators": 100}),
#         (MLPClassifier, {}),
        (AdaBoostClassifier, {}),
        (LDA, {"solver":"lsqr", "shrinkage":"auto"}),
#         (QDA, {}),
    ]


    model_time =  []
    n_splits = cv_splits
    columns = ["classifier", "split", "metric", "score", "oversampling", "use_pca"]
    results = pd.DataFrame(columns=columns)
    min_max_scaler = preprocessing.MinMaxScaler()
    resampler = ADASYN()
    pca = KernelPCA(pca_components, kernel=pca_kernel)
    for i, (model, params) in enumerate(models):
        name = model.__name__+str(params)
        start = time()
```

```
        for j in range(n_splits):
            split = np.arange(int(len(X)/n_splits * j), int(len(X)/n_splits * (j + 1)))
            X_train, y_train = np.delete(X, split, 0), np.delete(y, split, 0)
            X_test, y_test = X[split], y[split]
            if oversampling:
                X_train, y_train = resampler.fit_resample(X_train, y_train)
            X_train = min_max_scaler.fit_transform(X_train)
            X_test = min_max_scaler.transform(X_test)
            if use_pca:
                X_train = pca.fit_transform(X_train)
                X_test = pca.transform(X_test)
            model_ = model(**params)
            model_.fit(X_train, y_train)
            y_pred = model_.predict(X_test)
            for metric in metrics:
                results = results.append(
                    pd.DataFrame([[name, j, metric.__name__, metric(y_test, y_pred), oversampling,
use_pca]],
                                 columns=columns)
                )
        end = time()
        print(name)
        result = results[(results["classifier"]==name)]
        for k, metric in enumerate(metrics):
            metric_name = metric.__name__
            vals = result.loc[result["metric"]==metric_name, "score"]
            print("\t%s: %f+-%f" % (metric_name, vals.mean(), vals.std()))
        model_time.append((end - start))
    return results, model_time

import warnings
warnings.filterwarnings('ignore')
print("### Plain Classifiers ###")
plain, _ = classifiers(X.values, y.values, oversampling=False, use_pca=False, cv_splits=30)
print("\n### Oversampling Classifiers ###")
oversampled, _ = classifiers(X.values, y.values, oversampling=True, use_pca=False, cv_splits=30)
print("\n### PCA Classifiers ###")
decomp, _ = classifiers(X.values, y.values, oversampling=False, use_pca=True, cv_splits=30)
print("\n### PCA/Oversampling Classifiers ###")
combined, _ = classifiers(X.values, y.values, oversampling=True, use_pca=True, cv_splits=30)
```

 # GridSearch: Set parameters values for each classifier

 # For KNN

```
k_range = range(1,10)
parameters =  {
"MLP":[{'classifier__hidden_layer_sizes':[(200,),(300,),(400,),(500,)]
                }],
"KNN":[{"classifier__n_neighbors":k_range, "classifier__weights":['uniform',
'distance'],"classifier__n_jobs":[1]}],
"GB" :[{'classifier__learning_rate': [0.1,0.01], "classifier__n_estimators":[25,50,100,150]}],
"RF":[{'classifier__n_estimators':[10,30,50,100], 'classifier__max_features': ['auto','log2',None],
                'classifier__min_samples_leaf': [0.2,0.4,1]}],
```

```
"DTC": [{'classifier__max_depth': [None,1,5,10], 'classifier__min_samples_split':[2,4,6,8,10,12],
'classifier__min_samples_leaf':[1,2,3,4,5,6,7]
                    }],
"SVC": [{'classifier__C':[1.0,10.0], 'classifier__kernel':['linear','rbf', 'sigmoid']}],
"BC":[{'classifier__n_estimators':[10,50,100,120,140]}],
"XGB": [{'classifier__booster':['gbtree','dart'],}],
"EXT":[{
'classifier__n_estimators':[5,10,100]}],
"LG":[{'classifier__penalty':['l1','l2']}]
          }
```

 # List of Classifiers to

```
models = {"KNN": KNeighborsClassifier(),
"RF": RandomForestClassifier(),
"GB":GradientBoostingClassifier(),
"DTC":DecisionTreeClassifier(),
"MLP":MLPClassifier(),
"SVC":svm.SVC(),
"BC":BaggingClassifier(),
"XGB": XGBClassifier(),
"EXT":ExtraTreesClassifier(),
"LG":LogisticRegression(),

      }
```

Source code from data.py file

 # Import important packages

```
import pandas as pd
import numpy as np
from sklearn.pipeline import FeatureUnion
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
np.random.seed(0)
```

 # List of features

```
feature_columns = ['District','Rainfall', 'Temp_max', 'Temp_min', 'Temp_mean',
     'Temp_range', 'Humidity', 'Wind_Dir', 'WasteWater', 'PoliticalEconomy']
```

 # A function to load the dataset

```
def load_data(file_name):
"""
     data loading
     file_name: Show the name of file to load
"""
     df = pd.read_csv(file_name)
```

```
    return df
# Load test dataset

def load_test_data(file_name, label_name):
   feature = pd.read_csv(file_name)
   label = pd.read_csv(label_name)
   return feature, label

# Create pipeline to preprocess the dataset before training

def data_pipeline(numeric_feature, categorical_feature):
    numeric_transformer = Pipeline(steps=[('imputer', SimpleImputer(strategy='median')),
                ('scaler', StandardScaler())])
    categorical_transformer = Pipeline(steps=[('imputer', SimpleImputer(strategy='constant',
fill_value='missing')),
                ('onehot', OneHotEncoder(handle_unknown='ignore'))])
    preprocessor = ColumnTransformer(transformers=[('num', numeric_transformer,
numeric_feature),
            ('cat', categorical_transformer, categorical_feature)])
    return preprocessor
if __name__ == "__main__":
   df=pd.read_csv('../data/train_data.csv')
   categ_column=["District", "WasteWater", "Year"]
   numeric_column=["Humidity", "Wind_Dir", 'Temp_mean', 'Rainfall']
   df_pipeline=data_pipeline(numeric_column, categ_column)
    print(df.head())
```

Source code from ensemble.py file

```
# Import important packages

    from sklearn.externals import joblib
    from sklearn.ensemble import VotingClassifier
    from sklearn.ensemble import BaggingClassifier
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.ensemble import ExtraTreesClassifier

# Plot the important features

    imp_feat_rf = pd.Series(rforest.feature_importances_,
    index=X.columns).sort_values(ascending=False)
    imp_feat_rf.plot(kind='bar', title='Feature Importance with Random Forest',
figsize=(10,6),color='g')
    plt.ylabel('Feature Importance values')
    plt.subplots_adjust(bottom=0.25)

# Metrics

    from sklearn.metrics import balanced_accuracy_score
    from sklearn.metrics import f1_score, classification_report
    from sklearn.metrics import confusion_matrix
    from data import *
```

```
from visualize import *
from logs import log
logger = log("../logs/ensemble_log")
```

 # Function to train the models

```
def fit_model(df, class_names, model_name):
    categ_column = ["District", "WasteWater", "Year", "PoliticalEconomy"]
    numeric_column = ["Humidity", "Wind_Dir", 'Temp_mean', 'Rainfall']
    df_pipeline = data_pipeline(numeric_column, categ_column)
    y = df['LabResult']
    X = df.drop(['LabResult'], axis=1)
    logger.info("fitting model")
```

 # Voting classifier

```
    random_classifier = RandomForestClassifier(class_weight='balanced_subsample')
    bagging_classifier = BaggingClassifier()
    extra_classifier = ExtraTreesClassifier()
    classifiers = [
            ("rf", random_classifier),
            ("bc", bagging_classifier),
            ("ext", extra_classifier),
    ]
    voting_clf = VotingClassifier( estimators= classifiers,voting='soft', weights=[1,1,1])
```

 # Create a pipeline

```
    pipe = Pipeline(steps=[('preprocessor', df_pipeline),
                                                    ('classifier', voting_clf)])
    logger.info("Train {}".format("Voting Classifier"))
    pipe.fit(X, y)
    y_pred = pipe.predict(X)
    cm = confusion_matrix(y.values, y_pred)
    print('Confusion Matrix : \n', cm)
    total = sum(sum(cm))
    sensitivity1 = cm[0, 0] / (cm[0, 0] + cm[0, 1])
    logger.info('Sensitivity :{:.3f} '.format(sensitivity1))
    specificity1 = cm[1, 1] / (cm[1, 0] + cm[1, 1])
    logger.info('Specificity : {:.3f}'.format(specificity1))
    balanced_accuracy = balanced_accuracy_score(y.values, y_pred)
    logger.info("balanced Accuracy: {:.3f}".format(balanced_accuracy))
    f1_tra = f1_score(y.values, y_pred)
    logger.info("f1 score: {:.3f}".format(f1_tra))
    clf_report = classification_report(y.values, y_pred)
    print(clf_report)
    plot_confusion_matrix(y.values, y_pred, class_names, title=model_name + "-cm")
    plt.savefig("../figures/train_figures/{}_{}_cm.pdf".format(model_name, "results"),
bbox_inches="tight")
    plt.close()
    #save the models
    joblib.dump(pipe, '../models/{}--f1-{:.3f}.pkl'.format(model_name, f1_tra))
    logger.info("-------------------------------")
def main():
```

 # Load dataset

```
data =  load_data('../data/train_data.csv')
```

 # Name of the classes

```
classes = ['cholera', 'health']
model = "Voting Classifier "
fit_model(df = data, class_names = classes, model_name= model)
if __name__ == "__main__":
main()
```

Source code from learner.py file

 # Import important packages

```
from sklearn.model_selection import GridSearchCV
from sklearn import preprocessing
from sklearn.model_selection import StratifiedKFold
from sklearn.externals import joblib
```

#Metrics

```
from sklearn.metrics import balanced_accuracy_score
from sklearn.metrics import  f1_score, classification_report
from sklearn.metrics import confusion_matrix
from data import *
from classifiers import *
from visualize import *
from logs import log
```

 # Set a logger file

```
logger = log("../logs/train_logs")
```
 # Function to train the models

```
def fit_model(parameters = parameters, models = models ):
   # name of the classes
   class_names = ['cholera', 'health']
```

 # Load the dataset

```
df=load_data('../data/train_data.csv')
```

 # List of category featuers
```
categ_column=["District", "WasteWater", "Year", "PoliticalEconomy"]
```

 # List of numerical featuerse
```
numeric_column=["Humidity", "Wind_Dir", 'Temp_mean', 'Rainfall']
```

 # Load preprocessing methods

```
df_pipeline=data_pipeline(numeric_column, categ_column)
y=df['LabResult']
X = df.drop(['LabResult'],axis=1)
   logger.info("fitting model")
   for model_name, model in models.items():
      #creae a pipeline
      pipe = Pipeline(steps=[('preprocessor', df_pipeline),
                     ('classifier', model)])
      logger.info("Train {}".format(model_name))

      # Gridsearch for each classifier

      clf = GridSearchCV(pipe, parameters[model_name], cv=5)
      clf.fit(X, y)
      print(clf.best_params_)
      logger.info("Best parameters values:{}".format(clf.best_params_))
      y_pred = clf.predict(X)
      cm = confusion_matrix(y.values, y_pred)
      print('Confusion Matrix : \n', cm)
      total = sum(sum(cm))
      sensitivity1 = cm[0, 0] / (cm[0, 0] + cm[0, 1])
      logger.info('Sensitivity :{:.3f} '.format(sensitivity1))
      specificity1 = cm[1, 1] / (cm[1, 0] + cm[1, 1])
      logger.info('Specificity : {:.3f}'.format(specificity1))
      balanced_accuracy = balanced_accuracy_score(y.values, y_pred)
      logger.info("balanced Accuracy: {:.3f}".format(balanced_accuracy))
      f1_tra=f1_score(y.values, y_pred)
      logger.info("f1 score: {:.3f}".format(f1_tra))
      clf_report = classification_report(y.values, y_pred)
      print(clf_report)

      # Plot the confusion matrix

      plot_confusion_matrix(y.values, y_pred, class_names, title=model_name +"-cm")

      #save the confusion matrix
      plt.savefig("../figures/train_figures/{}_{}_cm.pdf".format(model_name,"results"),
bbox_inches="tight")
plt.close()

# Save the model
joblib.dump(clf.best_estimator_, '../models/{}-f1-{:.3f}.pkl'.format(model_name,f1_tra))
      logger.info("-------------------------------")
def main():
   fit_model(parameters = parameters, models = models)
if __name__ == "__main__":
   main()
```

Source code from logs.py file

```
import os
import logging

# A function to create and save logs in the log files

def log(file):
    if not os.path.isfile(file):
        open(file, "w+").close()
    console_logging_format = '%(levelname)s %(message)s'
    file_logging_format = '%(levelname)s: %(asctime)s: %(message)s'

# Configure logger

    logging.basicConfig(level=logging.INFO, format=console_logging_format)
    logger = logging.getLogger()

# Create a file handler for output file

    handler = logging.FileHandler(file)

# Set the logging level for log file

    handler.setLevel(logging.INFO)

# Create a logging format

    formatter = logging.Formatter(file_logging_format)
    handler.setFormatter(formatter)

# Add the handlers to the logger

    logger.addHandler(handler)
    return logger
```

Source code from testing.py file

```
# Import important packages

from sklearn import preprocessing
from sklearn.externals import joblib

# Metrics

from sklearn.metrics import balanced_accuracy_score
from sklearn.metrics import  f1_score, classification_report
from sklearn.metrics import confusion_matrix
from data import *
```

```
from visualize import *
from logs import log

# Set a logger file

logger = log("../logs/test_logs")

# Function to test the model performance

def test_model(model_name):

# Name of the classes

    class_names = ['cholera', 'health']

# Load the test data and label

    features , labels = load_test_data('../data/test_data.csv','../data/true_label.csv')

# Load the model

    clf = joblib.load("../models/{}.{}".format(model_name,"pkl"))
    categ_column = ["District", "WasteWater", "Year", "PoliticalEconomy"]
    numeric_column = ["Humidity", "Wind_Dir", 'Temp_mean', 'Rainfall']
    df_pipeline = data_pipeline(numeric_column, categ_column)

# Features = df_pipeline.fit_transform(features)

    df_pipeline = data_pipeline(numeric_column, categ_column)
    logger.info("Testing model")
    logger.info("model name:{}".format(model_name))
    y_pred = clf.predict(features)
    cm = confusion_matrix(labels, y_pred)
    print('Confusion Matrix : \n', cm)
    total = sum(sum(cm))
    sensitivity1 = cm[0, 0] / (cm[0, 0] + cm[0, 1])
    logger.info('Sensitivity :{:.3f} '.format(sensitivity1))
    specificity1 = cm[1, 1] / (cm[1, 0] + cm[1, 1])
    logger.info('Specificity : {:.3f}'.format(specificity1))
    balanced_accuracy = balanced_accuracy_score(labels, y_pred)
    logger.info("balanced Accuracy: {:.3f}".format(balanced_accuracy))
    f1_tra = f1_score(labels, y_pred)
    logger.info("f1 score: {:.3f}".format(f1_tra))
    clf_report = classification_report(labels, y_pred)
    print(clf_report)

# Plot the confusion matrix

    plot_confusion_matrix(labels, y_pred, class_names, title= model_name + "-cm")

# Save the confusion matrix

    plt.savefig("../figures/test_figures/{}_{}_cm.pdf".format(model_name, "results"),
bbox_inches="tight")
    plt.close()
```

```
    logger.info("-------------------------------")
def main():

# Name of the model you want to test

    model = "GB-f1-0.304"
    test_model(model_name= model)
if __name__ == "__main__":
    main()
```

Source code from visualize.py file

```
#import important packages
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
import numpy as np
import itertools

# set colors
dark_colors = ["#A51C30", "#808080",
            (0.8509803921568627, 0.37254901960784315, 0.00784313725490196),
            (0.4588235294117647, 0.4392156862745098, 0.7019607843137254),
            (0.9058823529411765, 0.1607843137254902, 0.5411764705882353),
            (0.4, 0.6509803921568628, 0.11764705882352941),
            (0.9019607843137255, 0.6705882352941176, 0.00784313725490196),
            (0.6509803921568628, 0.4627450980392157, 0.11372549019607843),
            (0.4, 0.4, 0.4)]
SPINE_COLOR = 'black'


# A function to draw confusion matrix plot
def plot_confusion_matrix(y_true, y_pred, classes,
                        title='Confusion matrix',
                        cmap=plt.cm.Blues, fig_num=None):
"""
    This function prints and plots the confusion matrix.
    Normalization can be applied by setting `normalize=True`.
"""

    if fig_num is not None:
        plt.subplot(2,2,fig_num)
    fmt =   'd'
    cm = confusion_matrix(y_true, y_pred)
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=90)
    plt.yticks(tick_marks, classes)
    plt.title("")

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, format(cm[i, j], fmt),
                horizontalalignment="center",
```

```
            color="white" if cm[i, j] > thresh else "black")
    plt.ylabel('True label')
    plt.xlabel('Predicted label')



# a function to save the confusion matrix plot
def savefig(filename, leg=None, format='.pdf', *args, **kwargs):
"""
    Save in PDF file with the given filename.
"""
    if leg:
        art=[leg]
        plt.savefig(filename + format, additional_artists=art, bbox_inches="tight", *args, **kwargs)
    else:
        plt.savefig(filename + format,  bbox_inches="tight", *args, **kwargs)
    plt.close()
```

Appendix 5: Figures Showing Results Obtained During the Research Study

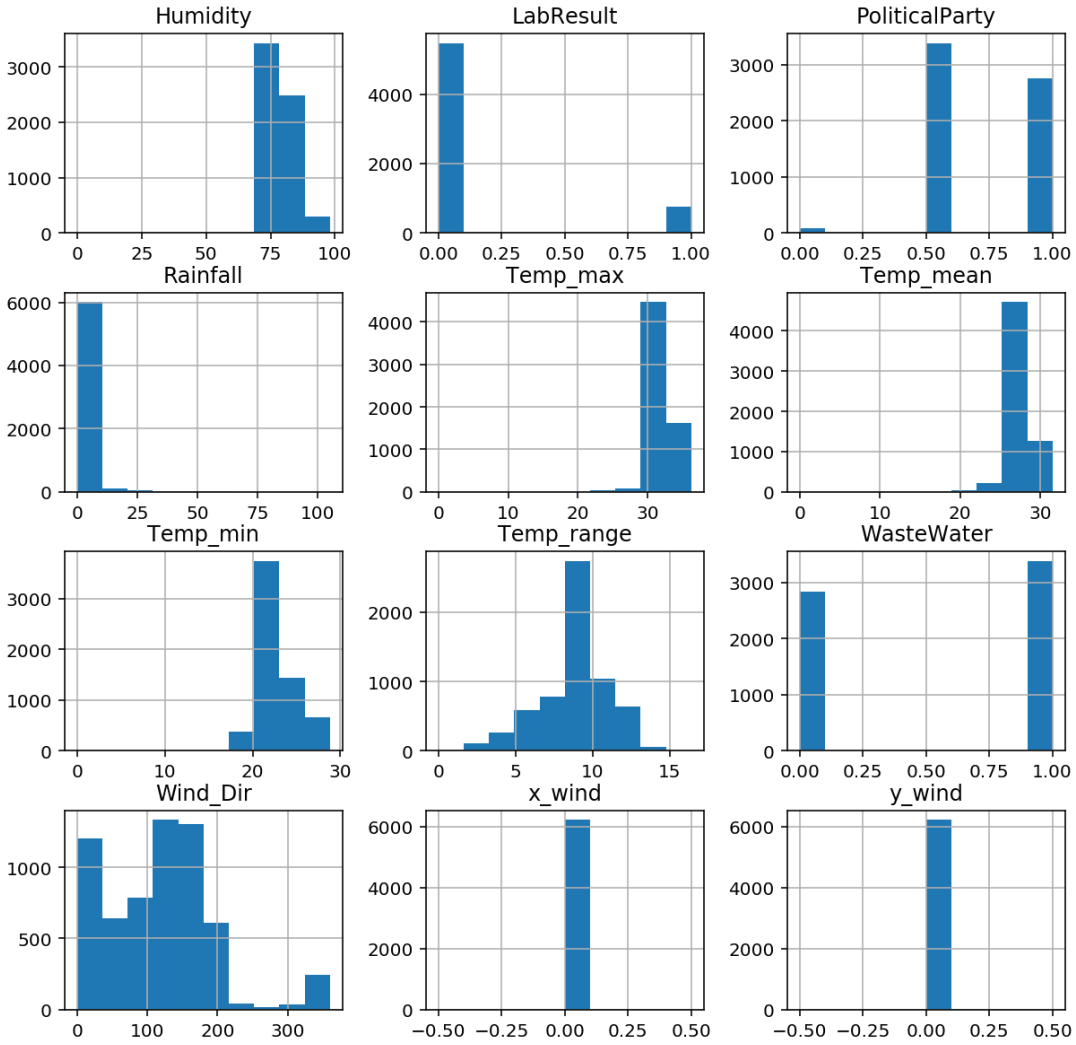


Figure 53: Variable in the Cholera Dataset

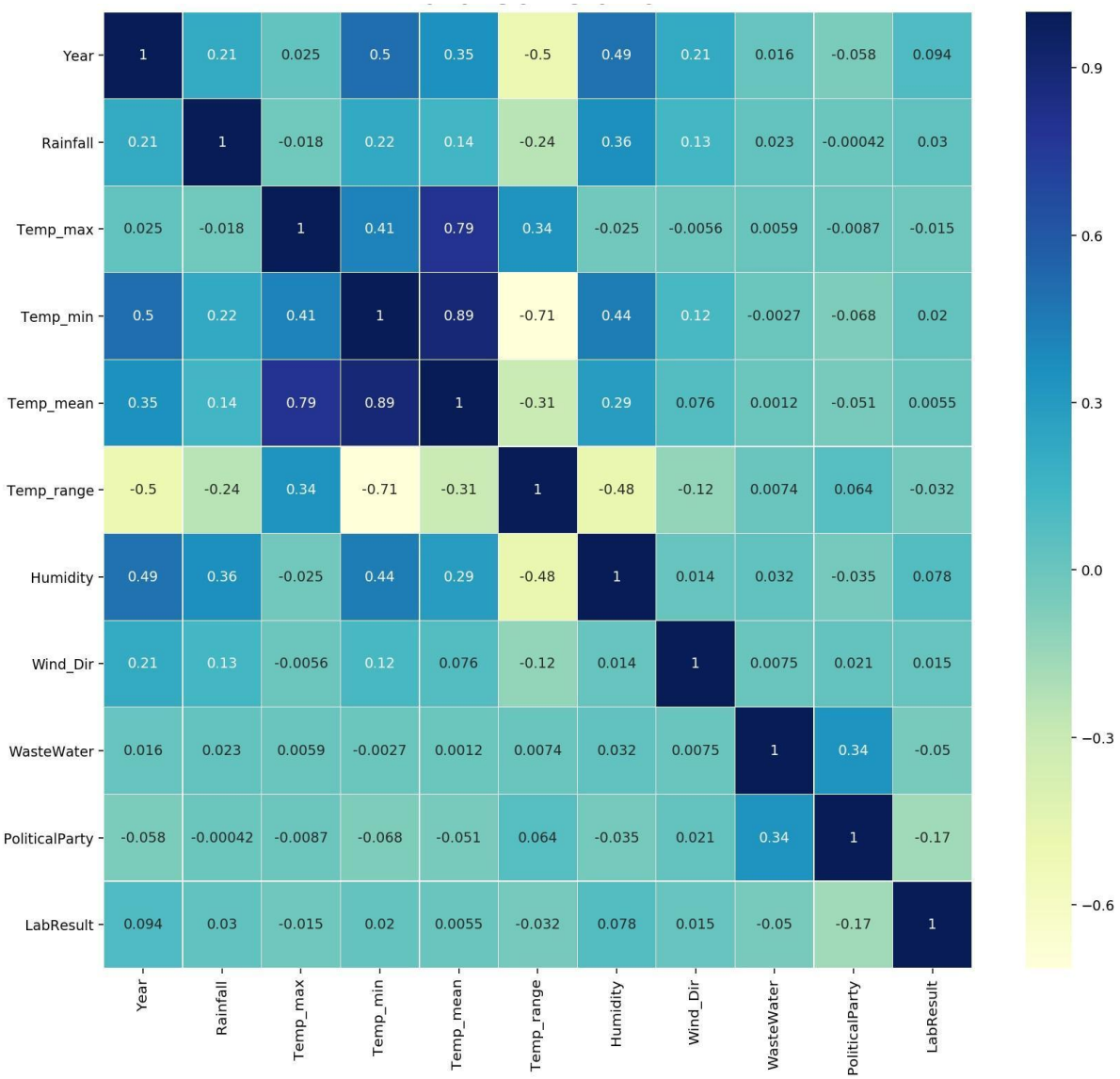


Figure 54: Data Correlation for Cholera Dataset